Generative Temporal Modelling of Neuroimaging - Decomposition and Nonparametric Testing

Hald, Ditte Høvenhoff

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Generative Temporal Modelling of Neuroimaging – Decomposition and Nonparametric Testing

Ditte Høvenhoff Hald

Kongens Lyngby 2016
IMM-PhD-2016-417

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk IMM-PhD-2016-417

Summary (English)

The goal of this thesis is to explore two improvements for functional magnetic resonance imaging (fMRI) analysis; namely our proposed decomposition method and an extension to the non-parametric testing framework. Analysis of fMRI allows researchers to investigate the functional processes of the brain, and provides insight into neuronal coupling during mental processes or tasks.

The decomposition method is a Gaussian process-based independent components analysis (GPICA), which incorporates a temporal dependency in the sources. A hierarchical model specification is used, featuring both instantaneous and convolutive mixing, and the inferred temporal patterns. Spatial maps are seen to capture smooth and localized stimuli-related components, and often identifiable noise components. The implementation is freely available as a GUI/SPM plugin, and we recommend using GPICA as an additional tool when performing ICA on fMRI data to investigate the effect of the temporal source prior.

In fMRI, statistical tests are used to investigate the significance of activation in specific brain regions. By extending the non-parametric testing framework to incorporate functional prior knowledge, an increase in sensitivity can be achieved, entailing better evaluations and conclusions. The functional prior knowledge is incorporated by use of a proposed Graph-Based Cluster Permutation Test (GBCPT), entailing the possibility to expand the use of cluster permutations to multiple applications, wherever a graph-based setup can be used.

Summary (Danish)

Formålet med denne afhandling er at undersøge to forbedringer for funktionel magnetisk resonans (fMRI) analyse; vores foreslåede dekompositionsmetode og en udvidelse til den ikke-parametriske testramme. Analyse af fMRI giver forskere mulighed for at undersøge funktionelle processer i hjernen, og kan dermed give indsigt i de neuronale koblinger der foregår under mentale processer.

Dekompositionsmetoden er en baseret på Gaussiske processer brugt som kildefordeling i uafhængig kilde analyse (GPICA), som derved tilfører en tidsmæssig afhængighed i kilderne. Der anvendes en hierarkisk model specifikation, der muliggør en specifikation der baserer sig både på en instantan og temporal foldnings mikstur version af kilderne. De afledte tidsmæssige mønstre og rumlige specificeringer identificerer glatte og lokaliserede stimuli komponenter, og giver ofte identificerbare støjkomponenter. Implementeringen er frit tilgængelig som en GUI/SPM plugin, og vi anbefaler at bruge GPICA som et supplerende redskab, til at udføre ICA på fMRI data, for at undersøge effekten af den tidsmæssige afhængighed i kilder.

I fMRI anvendes statistiske test til at undersøge betydningen af aktivering i specifikke hjerneregioner. Ved at udvide den ikke-parametrisk testramme til at inkorporere funktionel forhåndsviden, opnås en forøgelse af sensitiviteten, hvilket medfører bedre evalueringer og konklusioner. Det funktionelle forhåndskendskab er inkluderet i modellen ved brug af vores Graf Baserede Gruppe Permuterings Test (GBCPT), der medfører muligheden for at udvide brugen af gruppe permutaterings test til flere applikationer, hvor et graf baseret setup kan bruges.

Preface

This thesis was prepared at Cognitive Systems, Department for Applied Mathematics and Computer Science at Technical University of Denmark in fulfillment of the requirements for acquiring a PhD in Engineering. The work was fully funded by an awarded DTU scholarship, and conducted under guidance from the main supervisor professor Ole Winther, Department for Applied Mathematics and Computer Science, Technical University of Denmark. The work was carried out between April 2012 and June 2016.

The thesis deals with two topics of brain imaging analysis; a proposed independent component analysis algorithm and a graph based cluster permutation test. The thesis consists of a synopsis of the contributions; one chapter in the thesis, and one submitted article along with a companion implementation.

Lyngby, 20-June-2016

Ditte Høvenhoff Hald

Contributions

The thesis will focus on the contributions listed with Roman numbers below.

I Ditte H Hald, Ricardo Henao, and Ole Winther. Gaussian Process based Independent Analysis for Temporal Source Separation in fMRI. *Submitted to NeuroImage*, 2016.

II Ditte H Hald, Graph Based Cluster Permutation Tests, Chapter in Thesis (Article draft), 2016

Software

A GPICA Toolbox. With GUI for SPM or as stand-alone tool-kit, `https://github.com/dittehald/GPICA`

Acknowledgements

My years as a PhD student has developed me as a person and scientist, but it has also been a somewhat challenging, and not always easy to get through. So it is needless to say, that this would not have been possible without the help of a long list of people who have supported me on everything from thesis work to personal support. My greatest thanks to my supervisor professor Ole Winther. You have encouraged me to seek new knowledge and constantly develop my work, while guiding and supporting me through the process. I appreciate your feedback and open office door policy, especially for helping with new solutions whenever things where getting stuck.

In conjunction to Ole, I also want to thank Professor Lars Kai Hansen and PhD Ricardo Henao. Lars for proposing interesting projects and providing valuable guidance and project discussion; and Ricardo for his invaluable assistance in implementing the Gaussian Process based independent analysis.

I also owe a thanks to my co-workers and fellow students at Cognitive Systems DTU, without the supporting comments, and occasional social events, my PhD would not have been the same. A special thanks to my office mates PhD Camilla Birgitte Falk Jensen, PhD Laura Frølich, PhD student Kit Melissa Larsen and PhD Student Søren Føns Vind Nielsen who have given hard days a positive outlook.

Commas and spelling would not have been the same without my proof readers PhD Stine Harder, M.Sc. Martin Kristensson and GlobalDenmark. Thank you for your patience and due diligence.

Finally, the greatest thank goes to my way too patient man of my life and our wonderful daughter.

Contents

Chapter 1

Introduction

"All models are wrong, but some are useful", Box et al. [1987]

Exploration of the human brain and its functionality has fascinated scientists for ages, but the means to examine and investigate the brain's functionality have never been as good as they are now. In functional neuroimaging, the most established methods include Electroencephalography (EEG, scalp recordings of electrical potentials from the brain), Magnetoencephalography (MEG, measurement of magnetic fields generated by the brain's electrical potentials), and functional magnetic resonance imaging (fMRI, measurement of brain activity by registering minute changes in the magnetic field caused by changes in cerebral blood flow). Neuroimaging has a proven potential in connection with machine learning to obtain a non-invasive and non-ionizing view of the brain's functionality, see Figure 1.1.

Machine learning is a wide field, merging computational sciences, modeling and statistics that facilitate handling and processing of multivariate data, see Figure 1.2. Machine learning can mainly be classified into two categories; supervised, $p(y \mid x)$, or unsupervised learning, $p(x)$, based on whether both the input and output, x and y, or only the input data, x, is known. Supervised learning covers problems such as classification, regression and numerical analysis, while unsupervised learning deals with clustering and more generative factor modeling.

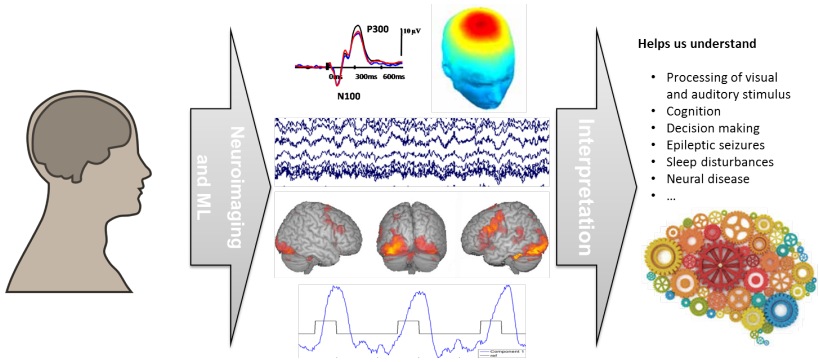


Figure 1.1: (Machine) Learning about the brain. Through neuroimaging, modeling and machine learning, deep insight into the brain can be gained. Illustrated here with EEG and fMRI visualizations. Knowledge about brain function can highlight the brain processing performed in conjunction with e.g. visual stimulation, and furthermore it can be used in diagnosis of neural diseases. Own figure design, subimages from [noa, Delorme and Makeig].

Unsupervised learning is said to "Extract an efficient internal representation of the statistical structure implicit in the inputs" Hinton and Sejnowski [1999], and makes it possibles to perform a more exploratory analysis.

Supervised fMRI data analysis is most commonly performed through Statistical Parametric Mapping (SPM8, Wellcome Trust Centre for Neuroimaging[1], a software plugin for MATLAB developed for data processing and hypothesis testing. SPM employs regression through the General Linear Model (GLM) to incorporate a stimuli function into the design matrix in order to create a linear model for the measured data. Based on the linear regression estimation of the effects and error estimate, a statistical measure can be used for hypothesis testing in order to detect significant stimuli-dependent areas in the brain. However, when trying to compare multiple voxel-time points in a statistical test, this method is subject to the multiple comparison problems (MCP).

Turning to unsupervised models, there are many applications for fMRI analysis. This thesis will focus on independent component analysis (ICA), since it is a great tool for performing exploratory analysis to extract components for further analysis. Independent components (IC's) can be used for decomposition into stimuli-dependent and non-stimuli-dependent components, and e.g. for data de-noising. When using ICA as a preprocessing step, one can couple supervised with unsupervised analysis by subsequently performing a non-parametric test-

[1] `http://www.fil.ion.ucl.ac.uk/spm/`

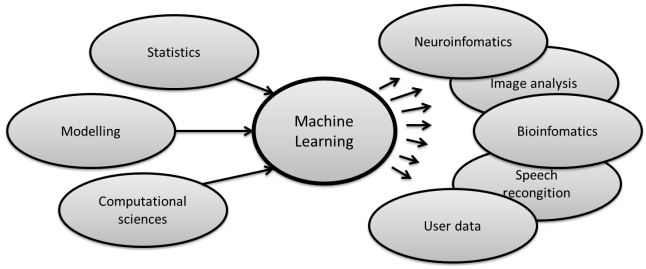


Figure 1.2: A view into aspects of machine learning. A great advantage of machine learning is the numerous multivariate data applications. This thesis will focus on neuroscience, especially fMRI.

ing. This leads to the two contributions of this thesis; a temporal generative model to perform ICA, and a graph-based cluster permutation test to mitigate the effects of MCP.

Outline of Contributions

The work of this PhD study has evolved around two main areas, both concerning two aspects of fMRI data analysis.

Contribution I: Gaussian Process based Independent Analysis for Temporal Source Separation in fMRI deals with the proposed method for temporal-based modelling of independent components through Gaussian processes (GP). The goal was to develop a pipeline for source separation in fMRI that featured temporal dependency. Whereas others have proposed GP based factor models for induced temporal dependency in the exploratory analysis, our contribution consists of a hierarchical model specification followed by inference of spatial source maps and temporal patterns by Markov Chain Monte Carlo, resulting in a direct fMRI application with a GUI and visual convergence diagnostic.

Contribution II: Graph-Based Cluster Permutation Tests propose a new extension to cluster permutation tests that features graph-based cluster connections to improve the statistical significance. Previously, only clustering based on time, space or frequency has been used, but graph-based clustering paves the way for multiple application possibilities, since networks can be specified on multiple bases, depending on what prior knowledge we wish to incorporate into the statistical analysis. In fMRI, the graphs incorporated were based on resting

state covaiance, and the preliminary work shows promising results on improving the test statistic.

Outline of the Thesis

This thesis is written as a synopsis, to focus on the contributions in Appendix A and Chapter 4.

Chapter 2 the thesis building block is briefly introduced, covering functional brain imaging, Independent Component Analysis, and Bayesian Inference.

Chapter 3 gives an introduction to the contribution of Gaussian-process-based independent analysis for temporal source separation in fMRI. The chapter covers a short introduction, additional results for elaboration, and further discussion and considerations in connection with the article.

Chapter 4 is a little untraditional as it serves as a contribution in itself, and should be read as a draft in preparation for future publication. This format has been chosen to make it possible to present the project work flow, leading to an open discussion on problems encountered and future modifications. It coves background, simulated and real data applications, through method description, and following findings and discussions.

Chapter 5 discusses the thesis' main results and relates them to each other and to some of the difficulties experienced in the field of neuroscience and machine learning.

Chapter 2

Preliminaries

The following will introduce some of the building blocks used in the thesis. The aim is to give a taste on some of the overall background, so please refer to the separate contributions for more topic specific details. This chapter will focus on three corner stones, namely functional brain imaging (Section 2.1), the Cocktail Party Problem (Section 2.2) - and how to solve it with Independent Component Analysis (Section 2.4), and finally a short intro will be given to Bayesian Inference (Section 2.4).

2.1 Functional Brain Imaging

One of the most interesting methods for functional brain imaging utilizes MRI. Functional MRI (fMRI) is of interest as it is a non-invasive, and non-ionizing technique, allowing researcher to asses neuronal activity in conjunction with stimulation (visual, auditoral etc.). It relies on the inherent coupling between neuronal activity and cerebral blood flow changes. Most fMRI techniques utilizes the blood oxygenation level dependent (BOLD) signal as a contrast agent. BOLD emerges from metabolic changes that occur in the local brain areas that are stimulated. The BOLD effect is based on the magnetic properties of hemaglobin (Hb) which, when oxygenated is diamagnetic, and when deoxygenated is paramagnetic. Neural stimuli, will consume glucose which requires oxygen to

break down, and therefore increase local consumption of oxygen in an area. This phenomenon is called the hemodynamic response (HDR), and can be modelled as seen in Figure 2.1. The use of oxygen to break down glucose, and the consumption of glucose, will both induce an increased blood flow and blood volume in the area. The resulting change of ratio between deoxygenated Hb (dHb) and oxygenated Hb, will cause a reduction of the magnetic field inhomogeneity and thereby increase the local MR signal slightly [Buxton et al., 2004]. The BOLD signal will increase approx. 2 seconds after initiated neuronal activity. If the neuronal activity continues, the signal will reach a plateau at approx. 5-8 seconds. Following brain activity, the BOLD response will perform an "undershoot", as oxygenated blood will flow slower than brain activity, causing the signal to drop below the baseline, before normalizing. Large subject to subject variance has been observed, while inter subject variances is lower, [Aguirre et al., 1998]. Due to the relative slow nature of the HDR, sampling times, or repetition time

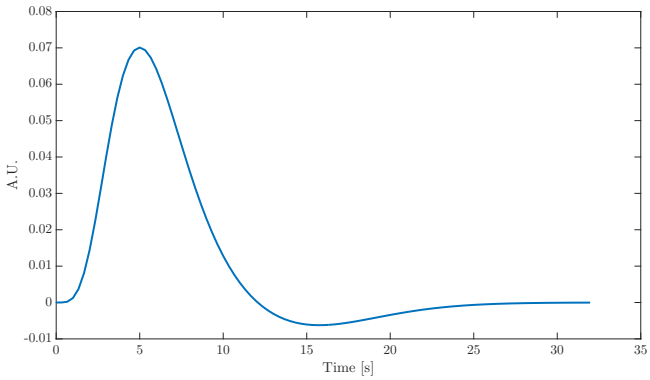


Figure 2.1: The HDR function with a TR = 0.333 s (as a dataset in this thesis), developed from SPM, for a delta

(TR), are typically in the order of 1-2 seconds. Compared to other neural activity measurement methods (EEG or MEG), this is a low temporal resolution, but on the other hand fMRI has the advantage of actually coupling neuronal activity with a spatial location, which makes it a attractive for neuroimaging.

2.2 The Cocktail Party Problem

2.2.1 Definition

The "cocktail party problem" (CPP) was first described in 1953 by Colin Cherry, as the human ability to recognize one person's speech among others interfering voices and sounds [Cherry, 1953]. The expression covers the human auditory systems ability to separate incoming signals, and attend to the one in focus. This type of scene segmentation is widely investigated in neuroscience – in both visual and auditory systems [Larsen et al., 2000]. It is of great interest to mimic the human perception, in regards to the ability of extracting source signals for countless of applications in machine learning. Extracting source signals from a set of mixed signals is known as "blind source separation (BSS) and can be seen as the "machine learning solution" to the cocktail party problem.

For more on the auditory perception regarding the cocktail party problem, see the thorough review by Haykin and Chen [2005]. McDermott [2009] elaborates on the two challenges of the CPP: sound segregation and directing attention.

2.2.2 Background

Blind source separation can be seen as an unsupervised decomposition method, and covers the process of finding hidden internal representations of the data, i.e. to decompose the data into its internal representations [Bishop, 2006]. Specifying the model with its sources, Z, the received mixed signal, X, and the mixing process, W:

$$X \approx WZ \tag{2.1}$$

For the CPP the sources are the speaker, persons and music at the party and the received mixed signal is represented as the microphones in Figure 2.2.

Blind source separation covers a wide range of algorithms with different conditions/constraints in order to overcome the under-determined aspect of the problem, and thereby narrow the set of possible solutions. Examples of methods for blind source separation include inter alia Principal Components Analysis (PCA) (though not cable to solving the CPP, since the representation of W and

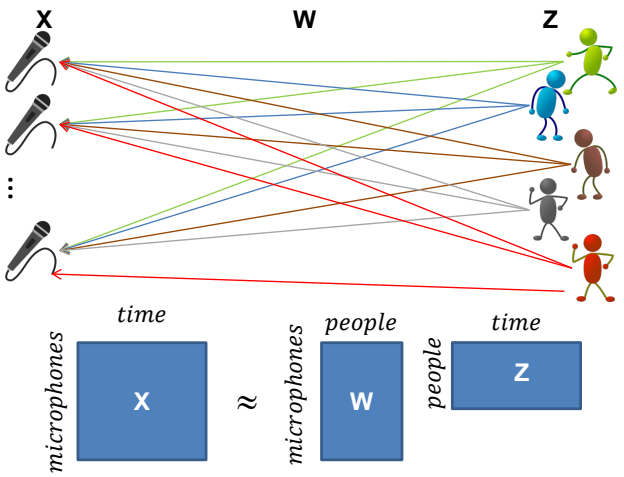


Figure 2.2: The cocktail party problem. Visualized with the partying people, Z, and the measurements units, the microphones, in X. The mixing matrix, W illustrated the sources mapping to the measurement places.

Z will not be unique)[1], Independent Component Analysis (ICA), Non-negative matrix factorization [Lee and Seung, 1999] and common spatial pattern [Koles et al., 1990].

BSS is not only applicable to the auditory cocktail party problem, but has a huge multidimensional applicability, covering inter alia (neuro)-imaging and musical decomposition. The following will focus on ICA, as it is one of the most used algorithms for solving BSS and thereby the CPP.

[1]In PCA the singular value decomposition is used: $X = U\Sigma V^\top$, where U and V are orthonormal, and Σ is diagonal. In the factor analysis/ICA view we can write $X = U\Sigma V^\top = U\left(\Sigma V^\top\right) = WZ = (UQ)\left(Q^{-1}\left(\Sigma V^\top\right)\right) = \hat{W}\hat{Z}$. The representation for W is thus not unique because we can always multiply by a rotation matrix Q from the the right $U \to UQ$. Like-wise Z is not unique because we can multiply by Q^{-1} from the left.

2.3 Independent Component Analysis

The classical model formulated for independent component analysis for P[2] components is:

$$X = WZ + \epsilon \tag{2.2}$$

$$X = \sum_{p=1}^{P} w_p z_p + \epsilon. \tag{2.3}$$

The components z_p are assumed to be mutually statically independent (uncorrelated) and non-Gaussian. Depending on the model, the error ϵ term can be excluded, but is defined to have zero mean and σ^2 variance. The ICA model is a generative model, since X can be generated by a process of mixing the latent variables z_p. ICA maximizes the statistical independence of the components, and some measures of independency include: Minimizing the mutual information (maximizing the log likelihood), as in *e.g.* the Infomax algorithm Bell and Sejnowski [1995]. Another method uses entropy/negentropy (since maximizing negentropy, minimizes the mutual information, as in fastICA Hyvärinen [1999]). The applications of ICA are manifold and covers neuroimaging, including fMRI and EEG, audio and hearing aids, speech recognition, music decomposition, images and text, [Larsen et al., 2000, Hastie et al., 2005]. For more background on ICA, please refer to *e.g.* the survey article by Hyvärinen and Oja [2000] or Hyvärinen et al. [2001], or the "original" ICA proposed by Comon [1994].

The following will shortly introduced the two well-established methods in the field:

The Molgedey-Schuster Independent Component Analysis (icaMS) The icaMS decorrelation algorithm, Molgedey and Schuster [1994], utilize that the mixing matrix, W, can be estimated using the covariance to find the quotient matrix. Eigenvalue-decomposition is performed on the quotient matrix to achieve W, since the quotient matrix has W as its eigenvectors. IcaMS requires no iterations, and no noise term is included in the model. The independent sources are expected to have different autocorrelation functions, [Molgedey and Schuster, 1994, Larsen et al., 2000], and the algorithm main advantage is to not require parameter tuning. An implementation of the algorithm is found from noa [2002], where Hansen et al. [2001] gave a likelihood specification of the problem.

The InfoMax based Independent Component Analysis Infomax ICA by Bell and Sejnowski [1995] minimize the Mutual Information between the

[2]Note when $p\,()$ is used, it symbolizes the probability, and not the p'th source

components and thereby create maximal independent components. Minimizing the MI is done through maximizing the log likelihood. With a noise free model. The model reads,

$$X = (X)_{i,j} = \sum_{p=1}^{P} W_{i,p} Z_{p,j}, \tag{2.4}$$

with i, j being measurements and time sample points, respectively.

The following will derive the log likelihood for the noise free ICA, from [MacKay, 2003] (with a bit more simplified notation). The likelihood function reads

$$p\left(X \mid W\right) = \int p\left(X \mid W, Z\right) p\left(Z\right) dZ, \tag{2.5}$$

where $p\left(X \mid W, Z\right) = \delta\left(X - WZ\right)$, since a noise free model is assumed, and the prior $p\left(Z\right)$ is the assumed source distribution. Thereby,

$$p\left(X \mid W\right) = \int \delta\left(X - WZ\right) p\left(Z\right) dZ. \tag{2.6}$$

By introducing the change of base, $\hat{X} = WZ \Leftrightarrow Z = W^{-1}\hat{X}$ and thereby $d\hat{X} = W dZ$. From here it follows that

$$p(\hat{X}) = \det\left(\frac{dZ}{d\hat{X}}\right) p\left(Z\right) \tag{2.7}$$

$$p(\hat{X}) = \frac{1}{|\det\, W|} p(W^{-1}\hat{X}). \tag{2.8}$$

The likelihood now reads,

$$p\left(X \mid W\right) = \frac{1}{|\det\, W|} \int \delta\left(X - \hat{X}\right) p\left(W^{-1}\hat{X}\right) d\hat{X}\ , \tag{2.9}$$

And by used of integrating delta functions the likelihood is found to be,

$$p\left(X \mid W\right) = \frac{1}{|\det\, W|} p(W^{-1}X), \tag{2.10}$$

and thereby the log likelihood can be calculated, and maximized with a maximization procedure of own choice.

2.4 Bayesian Modeling

In the Statistical learning cycle, [Gelman et al., 2014], visualized in Figure 2.3, the iterative learning process consists of three steps; Modeling, Inference and

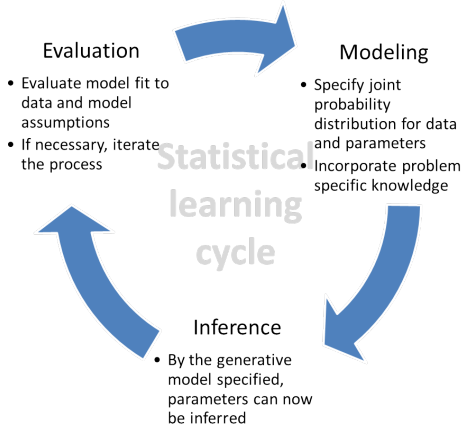


Figure 2.3: Statistical learning cycle, partly from [Gelman et al., 2014], where the tree corner stones are shortly described.

Evaluation. These will be the topic of following section, keeping the focus on relevant methods for this thesis.

The first step in the learning cycle describes the modeling of data in the Bayesian framework, consisting of specifying all assumptions and dependable variables in joint probabilities and prior distributions. By working with probabilities to describe the model, simple rules from probability theory apply. Bayes rule is the fundamental corner stone, specifying the relation between the posterior, the likelihood, the prior, and the marginal likelihood as in Eq. 2.12:

$$p\left(A \mid B\right) = \frac{p\left(B \mid A\right) p\left(A\right)}{p\left(B\right)}. \tag{2.11}$$

With a generative data model, with the data y, and the parameters θ, Bayes theorem consists of the probability of the data given the parameters $p(y \mid \theta)$, the probability distribution of the parameters, before observing the data, $p(\theta)$, and the marginal likelihood, $p(y) = \int p(y, \theta) d\theta$,

$$p\left(\theta \mid y\right) = \frac{p\left(y \mid \theta\right) p\left(\theta\right)}{p\left(y\right)}. \tag{2.12}$$

With Bayesian modeling it is possible to derive generative models for the specified models; but knowing that modeling usually cannot be performed precisely, inference models are used to approximate the posteriors and estimate unknown variables, [Bishop, 2006]. Multiple inference methods exist, (covered by e.g. Gelman et al. [2014], Bishop [2006]), but in this thesis the sampling is based on

Markov Chain Monte Carlo (MCMC). This stochastic technique approximates the expectations by a finite number of samples, and is suitable for complex model distributions, since it is constructed to have its equilibrium distribution at the target posterior distribution, [Gelman et al., 2014]. MCMC is random-walk based, and keeps a chain of the previous samples, so new samples can be drawn proposed by the previous. It is expected that each drawn sample becomes more and more likely to be from the posterior distribution. At convergence (after a warm-up period), the MCMC chain can be used as samples sampled from the target distribution in the further analysis, [Gelman et al., 2014].

Multiple MCMC algorithms exist, where two popular methods, Metropolis Hastings (MH) and Gibbs sampling (actually from the same family of MCMC methods), are especially relevant for this thesis. Gibbs sampling performs iterative sampling of conditionals, by sampling from one distribution while keeping the others stationary. MH is often used when Gibbs sampling cannot be performed e.g. when conjugate priors have been specified, and uses a proposal distribution to propose new samples, but only keeps the the candidate sample if it passes an acceptance/rejection rule, [Gelman et al., 2014].

Finalizing the statistical learning cycle, it is of great importance to evaluate the created model, and iterate the model specifications and the inference schemes accordingly, if the model and data do not match. Furthermore, for MCMC it is important to investigate convergence, since this limit is the sampling performed from the target distribution.

Chapter 3

Gaussian Process Based Independent Component Analysis

The following chapter is based on the article in Appendix A on *Gaussian process based independent analysis for temporal source separation in fMRI*. This chapter will focus on model elaboration and considerations regarding choices taken towards the end results reported in the article. The reader is therefore referred to Appendix A for further model details and results. This chapter will shortly introduce the work (Section 3.1), present additional results that were not included in the article (Section 3.2), then discuss model considerations (Section 3.3), and end comments in the conclusion (Section 3.4).

3.1 Model Summary

Gaussian Process based Independent Component Analysis (GPICA) is a proposed method to extract independent components in fMRI, as summarized in Figure 3.1. By the use of Gaussian process priors, it takes advantage of the incorporated temporal dependency to extract interpretable classifiable compo-

nents from the data. The novel contribution in this work covers the hierarchical model specification, holding both instantaneous and convolutive mixing, and the direct fMRI application toolbox, available here[1]. Figure 3.1 provides a concentrated view of the algorithm, covering the model setup, specifications and a short output demonstration.

There exist a decent amount of GP based regressing and GP incorporation into factor models [Yu et al., 2009, Luttinen and Ilin, 2009, Schmidt and Laurberg, 2008, Schmidt, 2009], and preliminary work on GP source separation has furthermore been formulated by Park and Choi [2007, 2008]. But for fMRI analysis we found it of great interest to build and explore a Gaussian process based source separation procedure, that can: feature temporal support for the sources, be extended to a convolutive model, and foster easily analyzable length scale parameters.

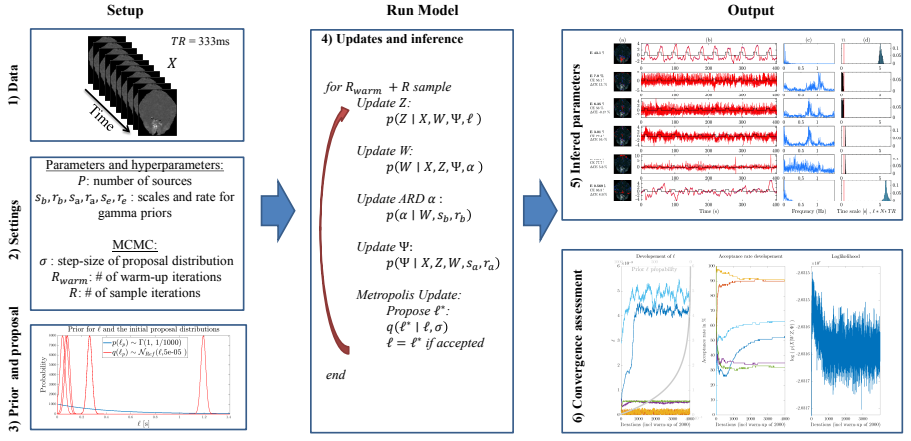


Figure 3.1: Schematic figure of GPICA, including model setup, concatenated updates, inference formulas and the common output for single-slice data. Setup includes 1) data and knowledge of TR. 2) Specifying parameters and MCMC settings. In 3) the prior for ℓ is visualized together with the initial proposed reflective Gaussian distributions with the specified step-size. 4) Example of the running model, with updates of the inference scheme and the Metropolis sampling. The model output covers 5) the inferred parameters, and 6) the convergence assessment. For the sake of simplicity, single-slice data is used for visualization in this schematic form. Taken from Appendix A.

[1] https://github.com/dittehald/GPICA

3.1.1 The Magic of Gaussian Processes

A Gaussian process (GP) can be seen as a generalization of a multivariate Gaussian distribution to infinitely many variables. Instead of being specified by multiple means and covariances, a GP for a process $f(t)$ can be fully specified by a mean function, $m(t)$, and a covariance function , $k(t, t')$, [Rasmussen and Williams, 2006].

$$f(t) \sim GP(m(t), k(t, t')). \tag{3.1}$$

Definition: "A Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions." [Rasmussen and Williams, 2006]

By the use of GP's, temporal dependency can be incorporated into the prior for the sought time sources. Instead of drawing samples from the posterior distribution for each element, the time series for each component is assumed to have its own distribution and thus is drawn from a Gaussian process. The dependency of previous elements is determined from the covariance function, where high correlation is assumed for the closest elements; and that the dependency will decrease exponentially in time, a kernel function is specified as the squared exponential covariance function [Rasmussen and Williams, 2006]. The length scale parameter ℓ gives the possibility to regulate the temporal dependency induced by the kernel. For further model specifications and model inference, please confer Appendix A.

3.1.2 Summarizing Model Results

In the article results on two data sets were presented; a decomposition of single slice data set and a full volume data set. The GPICA did produce a convincing source separation, with identifiable time source and associated localized spatial maps, even though no spatial assumptions were incorporated. The single slice data gave a clear demonstration of the interpretability of the length scales, entailing longer length scales the smoother the time source. The convolutive extension yields a great possibility to investigate the spatial prorogation of the source throughout the lags. The full volume data set was a bit more challenging, due to a high amount of low frequent drift and the considerable lower sampling frequency, but GPICA could still extract components that can serve a further basis of analysis. With GPICA a view into the convergence diagnostic was also opened, in order to observe the propagation and the samplers reliability.

3.2 Additional Results

As described in Contribution I, the results are acquired from the analysis of non-standardized data, honing a part of the goal - being as close to end-to-end [Saltzer et al., 1984] usage as possible. Even a correction to unit variance can change the data, by *e.g.* inflating more noise. But to simplify the parameter settings for the plug-in, it was decided to apply a mean-variance standardization to the datasets. By transforming the hyperparameters to account for the standardization, similar results as presented in the article are obtained. To give a short verification, the convolutive stimuli holding components is presented in Figure 3.2, and compares nicely to Appendix A Figure 6 (the rest of the findings compares as well, but are not presented here for brevity).

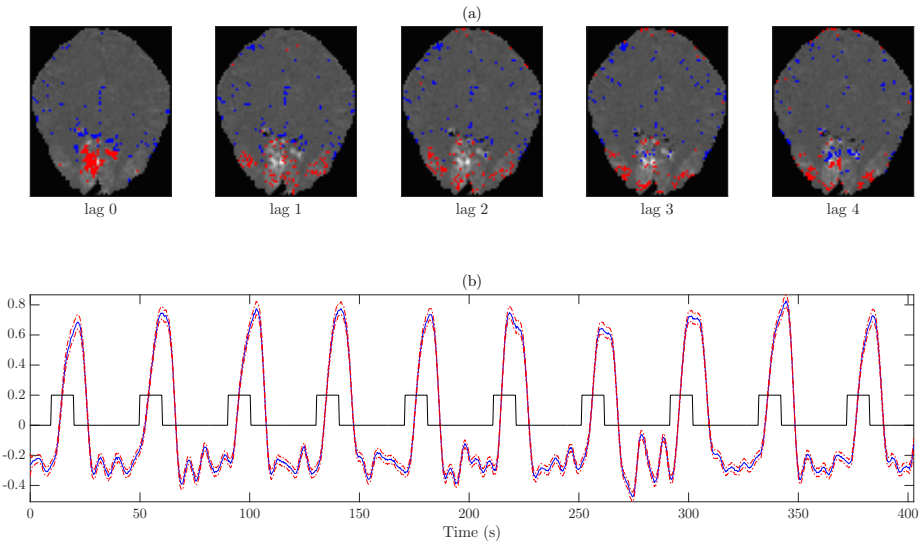


Figure 3.2: Convolutive results for the stimuli holding component. (a) Spatial map for all lags of the corresponding time series components in (b). The maps (a) show the 2.5% (blue) and 97.5% (red) quantiles of the mixing matrix's median, superimposed on an anatomical reference image created by averaging over all acquired images. In (b) the blue time source median is surrounded by the 5% and 95% quantiles in red.

Another interesting finding, not visualized for the 3D data in the article, is the results for the inferred precision parameters, α. In Figure 3.3a high variation areas in the components spatial maps are captured by the lowest 5% precision

element locations[2] . Note that the figure are produced by the end-sample iteration, since a all sample median procedure, as normally incorporated in the article, would hold a very data amount not suitable for storage. It has though been demonstrate, that the end and median sample results are similar if convergence has been reach after the warm-up period. The expected clear correspondence between the mixing matrix, W (Figure 3.3b), and its precision parameters testifies to a good spatial localization in the model. The low precision elements co-locates with the source spatial maps, since a low precision allows the parameter to be non-zero. These spatial maps (both α and W) can also be used for identifying noise/stimuli components (in connections with their respective time source), since *e.g.* a very unlocalized spatial source is more inclined to contain a noise component, than a more localized source.

A note on the visualization of the maps concerns the choice of visualization cut-off, *i.e.* when deciding to focus only on the 2.5% highest and lowest areas of the median. When investigating the distribution of the W values (as well as the 5% cut for α). But actually, the precision parameters could get a rough estimate on whether the W cut is set too high or too low. It also gives a clue about, that there are not always information in both the highest and the lowest areas, which could be considered when interpreting the results.

3.3 Model Considerations

There has been a great amount of decisions made throughout the model development, and with every decision follows a deselection of other opportunities. The following will cover a fraction of the model considerations that have been discussed during the model development. As with all models there are always room for improvement e.g. additional tests using simulated data in order to perform ground truth comparisons and error estimations are still to be performed. Improvements can furthermore be done in order to perform an estimation of the number of components to infer.

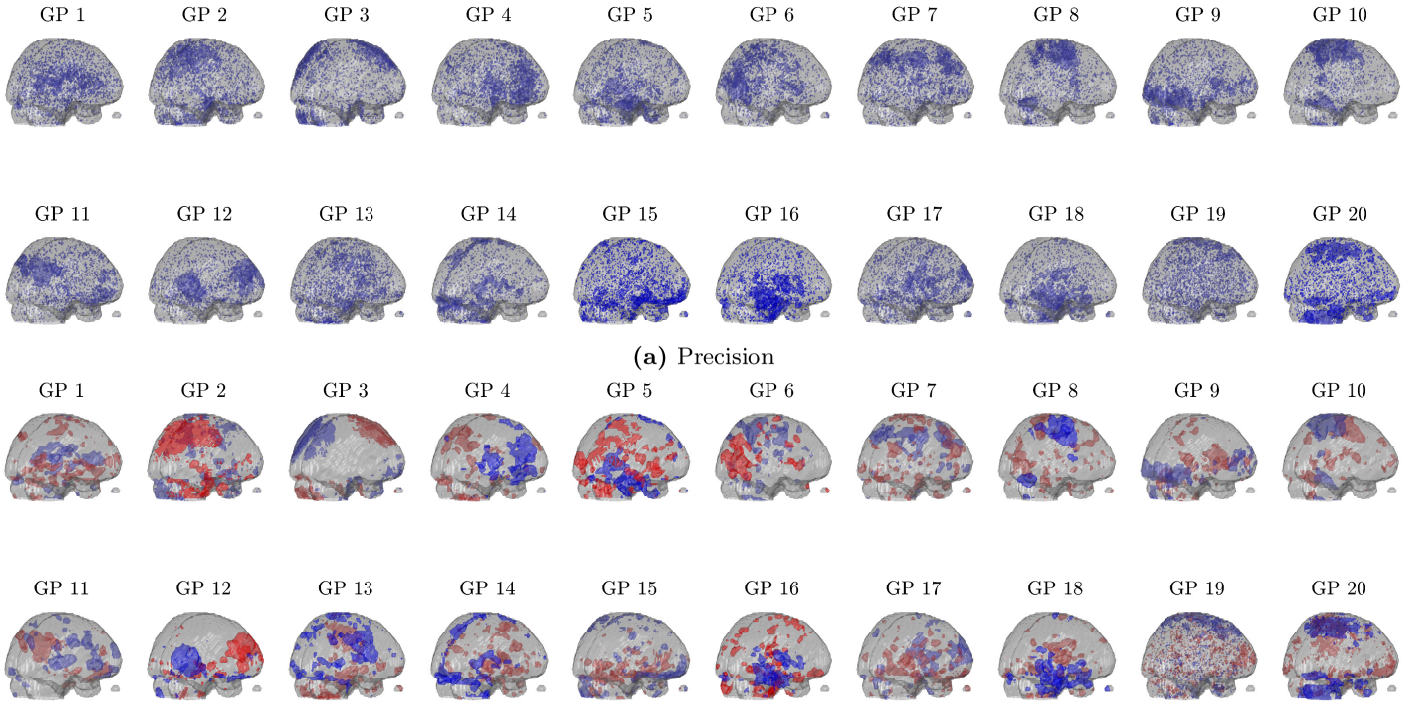
Convergence diagnostic Convergence diagnostic has been performed in a couple of different ways. Simple checks with multiple chains, *i.e.* different starting points, has testified to a stable performance, with converge to the same distribution. The visual inspection set-up designed for the plug-in, follows the convergence of the GP kernel hyperparameters. The potential scale reduction

[2] A small note on the 3D glass brains used throughout this thesis: The a bit nontraditional fMRI display tool is developed to give a fast and condensed 3D visualization. It requires a bit 'getting used to', but can in the toolbox been rotated and investigated as requested. The brain outline is extracted by the SPM mask.

[Gelman et al., 2014] was also investigated, (by running multiple chains, and splitting the after-warm-up samples in half, a number of sequences are acquired. By investigating within and between sequence variance, a so-called potential scale reduction, can explain the factor the estimated sequence distribution are scaled by, if the iterations are continued infinitely. No scaling factor, *i.e* potential scale reduction at 1, will indicate stationarity in the distribution [Gelman et al., 2014]. But since it is most correct to do convergence estimations on all inferred parameters, the amount of data were too large to incorporate into the toolbox. Thorough analysis to estimate a suitable number of iterative samples is a desirable future application.

Reporting Sampling Results When reporting sample results it was decided to report the median, since it is less sensitive to outliers (compared to the mean estimate), and gives a representation for all the collected samples (after the warm-up period). Other possibilities could be to report the end-sample result, and, as mentioned earlier, the end-sample would have a good correspondence to the mean and median results if convergence are reached after the warm-up.

Proposal Distribution For the GP kernel length-scale parameter distribution, a gamma has also been used as the proposal distribution. By specifying the variance to be dependent on the current sample value, the dynamic gamma function would adapt to have a high variance for larger values of ℓ, whereas the proposal distribution also could be used for fine tuning at the smaller values of ℓ. Thereby it were expected that the convergence would be reached faster, and that the estimation of ℓ would be more precise. But since no convergence enhancement were seen, we adapted the reflective Gaussian, and achieved a simple Metropolis update and the same convergence performance. Additional work on exploring different proposal distribution could be performed.



(a) Precision

(b) Spatial source maps

Figure 3.3: Correspondence between the precision elements, α and the spatial source maps. (a) End-sample results are presented for the inferred α parameter. Marked areas corresponds to high variance areas in the inferred spatial maps in (b). (Note that these results are generated from the standardized data, and the source numbers are slightly different than the ones given in the article).

3.3.1 Plug-in Specific Details

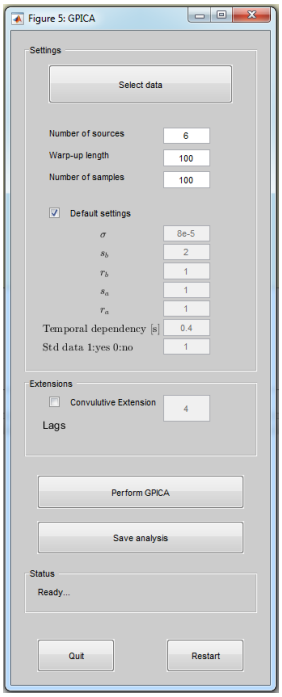


Figure 3.4: Screen shot of the GUI for the GPICA toolbox available from `https://github.com/dittehald/GPICA`.

The GUI for GPICA is depicted in Figure 3.4. It is possible to adjust all hyperparameters, or used the a default set shown to work acceptable on standardized fMRI data. In order to deliver a more smooth experience when running the model in the GUI, the convolutive model only reports the energy of the mean error, instead of calculating over all samples as reported in the article. Both energy calculation methods are available in the code package.

More work is currently being performed in order to speed up the model. The memory problem is, at the moment, a slightly issue when saving the data analysis, and when performing postprocessing of the data. Handling the huge W matrices for all samples requires a high memory server, and a reduction in the number of saved samples will make the software more applicable. A consideration would be to only save every 10th sample, and thereby still make it possible to report median sample results, without taking up too much memory.

3.4 Conclusion

A GPICA incorporating temporal source prior has been developed, with features extraction of smooth reliable sources, and appertaining localized spatial maps. The temporal dependency seen in neural data can thereby be supported in the inference of the sources, and the applications to other neuroimaging methods, like EEG, are therefore oblivious.

Chapter 4

Graph-Based Cluster Permutation Tests

4.1 Prolog

Inspired by the non-parametric statistical test performed by Maris and Oostenveld [2007] and Bullmore et al. [1999], and in connection with the graph-based-clusters approach suggested by professor L. K. Hansen, we propose a graph-based cluster permutation test for fMRI data.

The following is a summary of the work conducted with Professors Ole Winther and Lars Kai Hansen. Text, figures and large parts of the implementation were performed by the author, and the fMRI preprocessing was kindly performed by PhD Kasper Winther Andersen. The following is work in progress, and this chapter will introduce our work and touch upon some of the difficulties we encountered. It should be noted that the following is a proof of concept of a graph-based cluster permutation test, and functional brain connectivity is not the focus of the work. However, simple functional brain networks will be covered briefly, and wider applications discussed. The notation has been changed slightly compared to that used in Chapter 3.

4.2 Background and Introduction

A cluster-based permutation test is a non-parametric statistical test that deals with the Multiple Comparison Problem (MCP). The MCP arises when it is necessary to compare a large number of test pairs *e.g.* when evaluating the test statistic of a large number of time-voxel pairs for fMRI signals.

Bullmore et al. Bullmore et al. [1999] validated the use of a permutation test at cluster level in structural MRI data. Furthermore, Maris and Oostenveld [2007] used non-parametric testing for EEG and MEG data by clustering in the time and the frequency domain. Cluster permutation tests (CPTs) are also incorporated in FieldTrip[1] for EEG and MEG data, and in SnPM[2] for fMRI. The main advantage of these permutation tests is the freedom to choose any test statistic as a base. This allows for incorporation of prior knowledge and constraints to priors relevant for the biophysics behind the experiment, which will improve the sensitivity of the statistical test.

For this reason, clustering based on fMRI network graphs was proposed under the name the "Graph-Based Cluster Permutation Test (GBCPT)". By creating both functional and geometric graphs, it is possible to compare the significant areas identified, and show that a functional graph incorporation will increase the test statistic. The novel contribution consists of specifying graph-based clustering for permutation tests, that can not only be used in fMRI, but in a wide range of applications.

4.3 Theoretical Background

4.3.1 Statistics: t-test in SPM

fMRI data are usually (especially through SMP) modeled using General Linear Models (GLM):

$$Y = X\beta + \epsilon, \tag{4.1}$$

where Y is the fMRI signal (including the BOLD signal for activated areas), X is the design matrix holding the regressors expressing the model for the measured signal, ϵ describes the residual noise, which is assumed to be normally distributed

[1] `http://www.fieldtriptoolbox.org/` [Oostenveld et al., 2011]

[2] `http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/software/snpm` [Nichols and Holmes, 2001]

with zero mean and covariance σ^2 (each observation j having $\epsilon_j \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$). In the most simple setup, the design matrix holds the stimuli function convolved with the hemodynamic response and a constant factor (all known influential factors can be included in the design matrix). [Frackowiak et al., 2003]

In order to create the mapping of the activated areas, a null-hypothesis for each voxel is specified. The null-hypothesis states that the signal in a voxel is not affected by the stimuli, *i.e* the stimuli-related β value(s) is/are zero. Since the design matrix can hold multiple stimuli (*e.g.* for left and right hand), a contrast vector, c, is used to specify the test of the null-hypothesis, *i.e.* which regressor is to be tested. The null-hypothesis can stated as

$$c^\top \beta = 0, \tag{4.2}$$

where β values are estimated by minimizing the residual,

$$\hat{\beta} = \arg\min_\beta \|Y - X\beta\|^2 \tag{4.3}$$

and found using the pseudo-inverse,

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y \ . \tag{4.4}$$

A full rank design matrix will result in normally distributed $\hat{\beta}$ values, (by use of the minimum-variance unbiased estimator, $E\left[\hat{\beta}\right] = \beta$, and $\text{var}\left(\hat{\beta}\right) = \sigma^2 \left(X^\top X\right)^{-1}$). For the null-hypothesis, $c^\top \beta = 0$, it therefore follows that

$$c^\top \hat{\beta} \sim \mathcal{N}\left(c^\top \hat{\beta}, \sigma^2 c^\top \left(X^\top X\right)^{-1} c\right), \tag{4.5}$$

and the t-statistic ($t = \frac{\text{contrast}}{\text{normalized noise term}}$) can now be carried out as [Ashby, 2011, Frackowiak et al., 2003]:

$$t = \frac{c^\top \hat{\beta}}{\sqrt{\sigma^2 c^\top \left(X^\top X\right)^{-1} c}} \ . \tag{4.6}$$

4.3.1.1 The Multiple Comparison Problem

From the resulting summarizing t-statistic (the t-map of the entire data), a threshold must be applied in order to determine which voxels are significant. In order to do this, a problem occurs when it is necessary to compare multiple voxel-time points. When performing hypothesis testing, two type of errors can emerge, as described in Table 4.1. When performing multiple tests, the Type I

errors become the family-wise error rate, since the H_0 hypothesis now is a family-wise hypothesis (claiming that there is zero activation everywhere). fMRI gives rise to a massive MCP due to the high number of voxels (often >100,000) and the many time samples (>200). Given a small brain volume of 20x20x25 voxels (10,000 voxels), a p-value of 0.05 would still yield 500 false-positive voxels that by chance could cluster and be mistaken for significant clusters.

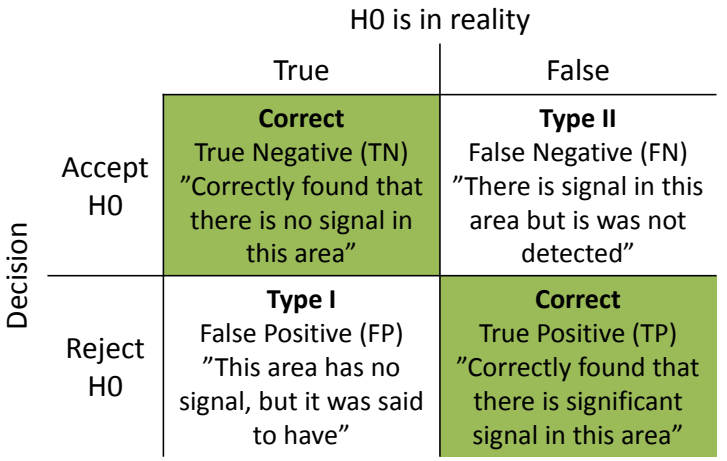


Figure 4.1: Types of errors in univariate hypothesis testing, also known as the confusion matrix or the contingency table. Note that the probability of making Type I errors is the level of significance, α. Adapted from Kohavi and Provost [1998], Freund and Johnson [2011] .

The standard SPM procedure overcomes the MCP by performing Random Field Theory (RFT) corrections, cluster level inference, or False Discovery Rate (FDR), [Frackowiak et al., 2003, Brett et al., 2004, Pernet and Pernet, 2016, Chumbley and Friston, 2009], but as mentioned, cluster permutation tests can be also be applied, [Nichols and Holmes, 2001].

4.3.2 Nonparametric Statistical Testing

Nonparametric statistical testing is a procedure to overcome the MCP. It is applied by constructing a histogram to achieve a non-parametric estimate of the probability distribution. The use of the permutation test to estimate the distribution is well-established [Good, 2013], since a permuted data test is expected to resenble the unpermuted version under the H_0 hypothesis. From rearranging the labels of the test, it is possible to achieve a permutation histogram as an

estimate of the permutation distribution (of the test statistic), and thereby assess p-values for rejection of the H_0 hypothesis, see Figure 4.2. The permutation test enables the freedom to choose any test statistic, thereby incorporating prior information *e.g.* clustering in time (EEG) [Maris and Oostenveld, 2007] and/or space (fMRI) [Bullmore et al., 1999, Nichols and Holmes, 2001], or by using networks to specify the neighborhood in which the t-statistics can be summed. As in Bullmore et al. [1999], Nichols and Holmes [2001], a cluster statistic is used as the test statistic, where the cluster mass is the summed voxel test statistic, t_m,

$$t_m = \sum_{i=1}^{v_m} t_i, \tag{4.7}$$

for M clusters with v_m voxels/pixels in each cluster.

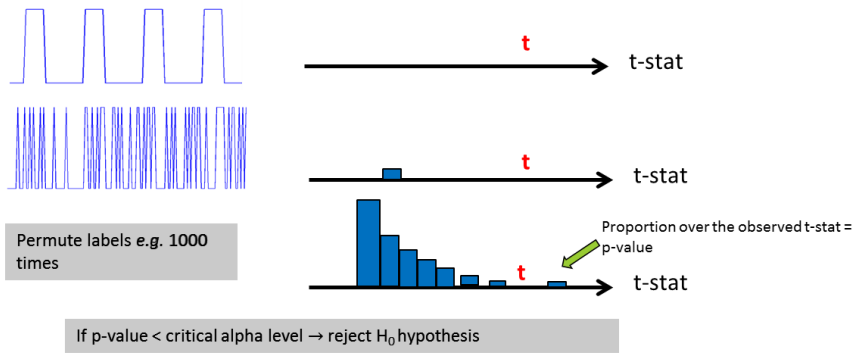


Figure 4.2: The idea behind permutation tests. Under a paradigm, the t-statistics are calculated (indicated with the red t). By permuting the paradigm, a new t-value is calculated, which will represent one observation in the permutation histogram (blue mark). After performing a number of permutations, the permutation histogram will estimate the permutation distribution, and the p-value for H_0 can thereby be assessed as the fraction of permutations above the observed t-value.

4.3.3 Graph-Based Modeling

A graph, or network, is a representation of linked pairs of objects, and a graph, $G = (V, E)$, can be defined by its objects, called nodes or vertices, V, and the links, called edges, between them, E. When dealing with images, the nodes are usually voxels of the image, or resampled versions of those, and the edges

are used to link different voxels together. Graphs can either be un-directed or directed, specifying if a link between two nodes is unidirectional or not. A graph's so-called adjacency matrix, A, is used to specify these links through a binary $A_{ij} = 1$, if a link goes from vertex i to vertex j. A so-called un-directed graph will always have $A_{ij} = 1, A_{ji} = 1$, and will therefore always be symmetric, $A = A^\top$. An example of a graph and its adjacency matrix is visualized in Figure 4.3

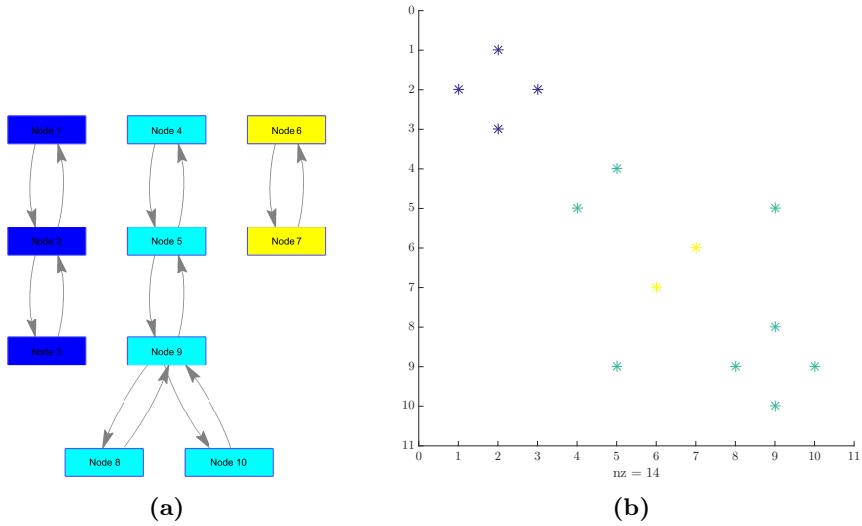


Figure 4.3: Illustration of graph connections in a symmetric graph with the vertices, $V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and the undirected edges, $E_{undir} = \{(1, 2), (2, 3), (5, 4), (9, 5), (7, 6), (9, 8), (9, 10)\}$. (a) is a graphic representation of the graph given by the adjacency matrix specified in (b). Note that the adjacency is colored according to the nodes' affiliation (and not according to the nodes' values, which in this case are always unity). The clustering of the graph components has been performed with Depth First Search [Tarjan, 1972].

Human Brain Networks

In fMRI it is of great interest to establish brain networks that can improve our understanding of how the human brain is connected. Resting state fMRI is of particularly great interest, since it has been suggested that the BOLD activation under rest will correspond to functionally relevant networks in the

brain, [Biswal et al., 1995, Damoiseaux et al., 2006, Raichle et al., 2001]. These complex networks can be estimated by investigating correlation of the temporal dynamics between voxels, or regions of voxels, [van den Heuvel et al., 2008, Biswal et al., 1995].

4.3.4 Connected Components

A way to retrieve graph information is to look for connected components in the network. Since only undirected graphs are in focus, it is only necessary to perform the search for connected components on the lower triangle of the adjacency matrix. The proposed GBCPT uses the 'Depth first search' algorithm by Tarjan [1972] from the Matlab standard graph toolbox to extract the connected components as shown in Figure 4.3 (a). For undirected graphs, the 'Depth first search' is shortly described with the graph in Figure 4.3 as an example. A random node is chosen and all connected nodes are visited and kept track off by listing them in a 'sequence' - to track the search stage a 'stack' is used to specify the current standing. As an example, choose node 4. From there a link makes it possible to visit node 5. From here it simply follows to visit node 9, and the sequence is now 4, 5, 9. Two unvisited nodes are now possible to visit, *e.g.* let it be node 8 (Stack: 4-5-9-8). From there it is only possible to go back to a node already in the sequence (node 9), and node 8 is therefore left, resulting in a stack specifying 4-5-9. An unvisited node can now be visited, and the sequence/stack becomes: 4, 5, 9, 8, 10 / 4-5-9-10. Now, no unvisited nodes are left and we move back through our track (to check for other branches/unvisited nodes not present in this example): 4-5-9-10 / 4-5-9 / 4-5 / 4. With an empty stack, it is guaranteed that a connected component is fully specified by the sequence 4, 5, 9, 8, 10, and a random node, not in the sequence, can now be visited and the process can proceed to specify more connected components.

4.4 Data Description

Two data sets were used to evaluate the developed method. The simulated data was used for the GBCPT proof-of-concept test, and the real fMRI data was used to test the GBCPT on real-life data, with noise artifact etc. The two datasets are described in detail below.

4.4.1 Simulated data

A simulated 3D dataset was created in order to prove the concept of GBCPT. The dataset is illustrated in Figure 4.4 and the following will describe the generation of the dataset. The full data volume is $25 \times 25 \times 32$ voxels holding a 'brain' masked to contain $20 \times 20 \times 25$ voxels. Two sources are incorporated; both being located with two activation spots in diagonal corners, as seen in (a). The blue and green sources are constructed from an activation pattern, as depicted in black, in subplot (b), convolved with the HDR function[3] constructed with a TR of 0.333 s. The resulting response to the stimuli function is shown in blue in (b), and a downscaled version in green. This is constructed in order to simulate an fMRI response in two spatially separated areas, with the same stimuli but with different amplitudes. The red source in (a) is a sine function with a lowered amplitude (0.8), as can be seen in red in (b). Random noise is then added to the data and the 3D matrix is filtered/convolved with a uniform $3 \times 3 \times 3$ filter. A resulting data slice is seen in (c) (the 3D location is illustrated by the green slice in (a)). The slice is extracted in the time instance illustrated by the gray line in (d), where two time series from each of the source areas are also illustrated (voxel location specified in (c) with the associated colors).

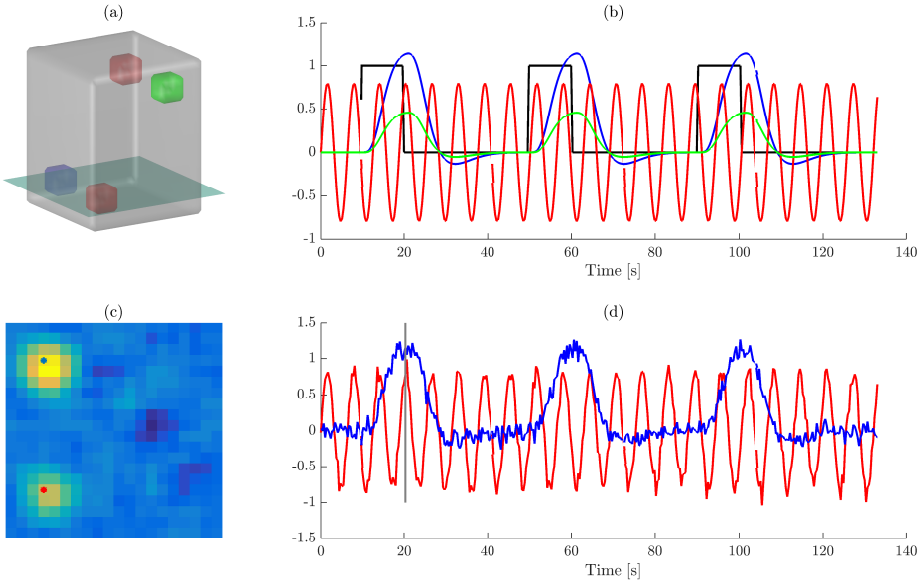


Figure 4.4: The simulated data.

In order to generate the resting state data, the procedure described above is

[3]From SPM `http://www.fil.ion.ucl.ac.uk/spm/`.

followed, and afterwards the signal strength is reduced, and noise is added. This simple setup is used, since the assumption regarding resting state data is that connected areas in the brain will convey, even when resting. In proving the concept for GBCPT, a relatively high correlation between the resting state and the active state is accepted.

4.4.2 fMRI Data

The data consists of an fMRI dataset[4] collected at Hvidovre Hospital from an experiment where the 28 test subjects were instructed to follow a finger tapping paradigm. The data was collected with a scan repeat time of TR = 2.49 s over 240 samples. The stimuli changed, alternating between left and right hand and starting with the left hand movement. Each active period was 20 s, followed by 10 s rest. During the resting period a fixation cross was shown in the middle of the screen; and there was a visual cue in the left and right conditions (red/green blinking dot at 1 Hz) to pace movements. The full data volume is $53 \times 63 \times 46$ voxels, but is masked to contain 60,678/153,594 voxels holding the brain volume. A resting state measurement was collected prior to the stimuli. The subject was instructed to relax with closed eyes during the measurement over 480 samples [Rasmussen et al., 2012].

The resting state data will be used for prior graph creation. The data under stimuli will be used for testing with the general linear model.

4.5 Method and Code Description

The following will summarize and serve as a guide for the GPCPT. All steps will be covered, so to keep track of the process, please refer to the graphical illustration in Figure 4.5. The letters and numbers of the subplots are used as a reference in the text. Stage 1 and stage 2 data is the resting state and the active state, respectively. The term 'prior graph' is used to specify the networks needed to be specified prior to the CPT. Note that the figure is constructed with data from the finger tapping experiment.

Framework for the network-based Cluster Permutation Test:

[4]Note same dataset as in Contribution I

Construct a functional 'prior graph' from a geometric graph specifying voxel neighbors under the active state (a.1), and a resting state covariance (a.2).

Find the t-maps (b) and extract sub-graphs with activation over a threshold value for both the functional and geometric prior graphs.

Find connected areas in the graphs and calculate aggregate t-values (c).

Perform a permutation test (permute the activation), and for each permutation, calculate the t-values (d).

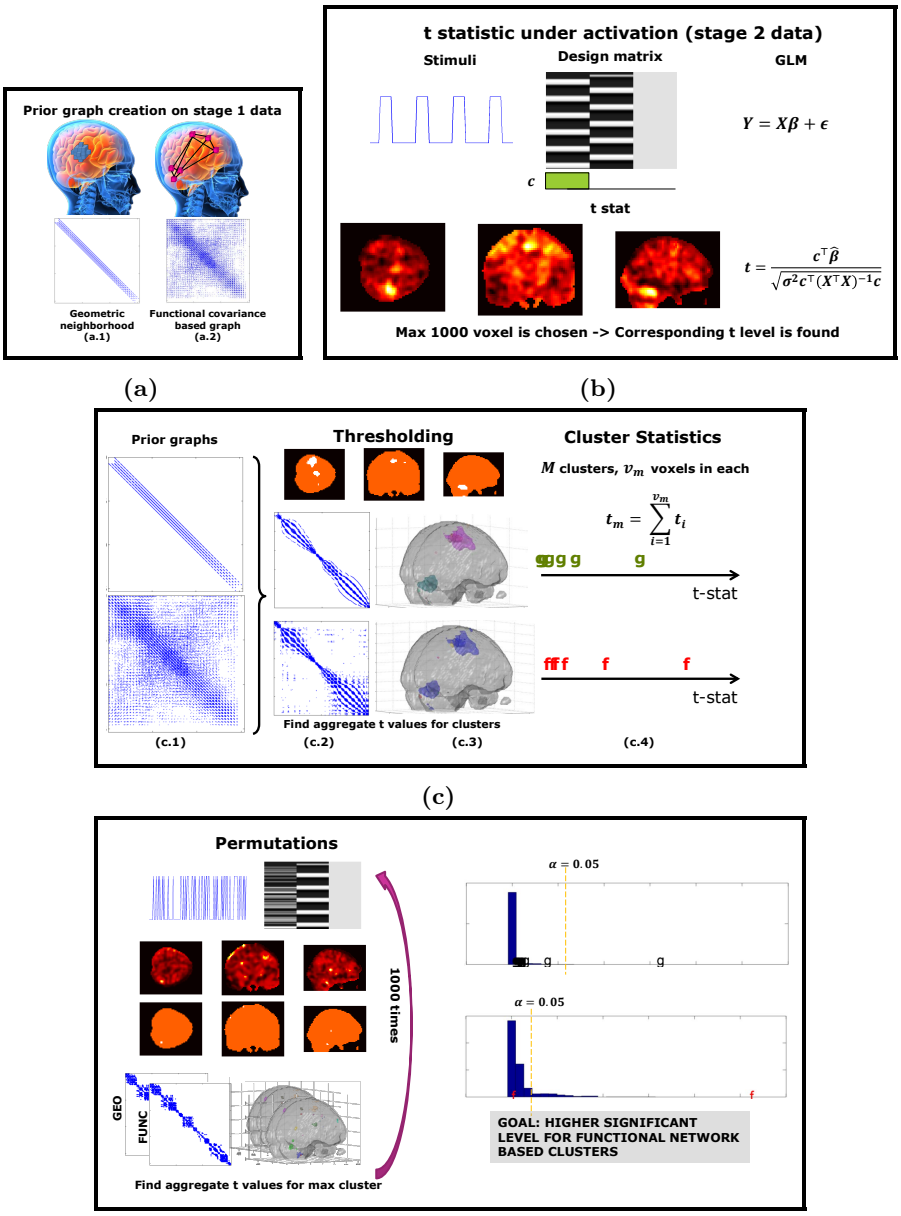


Figure 4.5: Step-by-step function illustration of the GBCPT. Please refer to the text for a thorough explanation. The functional and geometric prior graphs are calculated in the resting state (a), followed by t-statistic calculations during activity (b). In (c), the calculated t-statistics are thresholded to derive areas of activity. (d) Permutation is performed with new design matrices to assess significance of identified t-statistics.

(a.1) Geometric Graph Construction in Resting State

The geometric graph in (a.1) is created using a specified neighborhood around each voxel. Different geometric neighbors have been used, and for 3D data the number of neighboring voxels is: 6, 24, 32, 64, as specified by Figure 4.6. Note that many other neighborhood set-ups can be used (*e.g.* a 3 by 3 cubic set-up). The following were chosen to give a wide spread of the number of neighbors tested against the functional graph set-up.

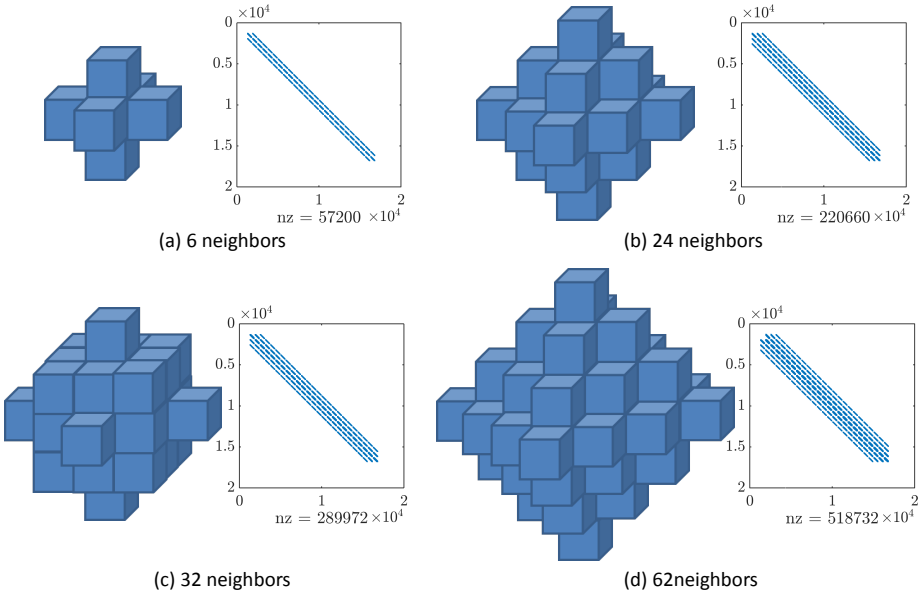


Figure 4.6: Geometrical neighbors in 3D together with the associated adjacency matrices for the simulated data in Figure 4.4. Note that the graphs have been constructed for the entire data volume (the unmasked data), and therefore more sparse.

(a.2) Functional Data-Based Graph Construction in Resting State

The subject-specific functional graph in (a.2) is created using the covariance matrix as in Eq. 4.8.

$$C = \frac{X^\top X}{N-1}, \tag{4.8}$$

where X is the data matrix with zero mean. Note that the graph construction is performed on non-smoothed data (normally a preprocessing step in the SPM pipeline), in order not to capture inflicted smoothing on the graph.

To be able to build the functional adjacency matrix for a full fMRI resting state dataset, the computation of the covariance matrix is split into blocks (otherwise it would cause a memory problem). The number of highest correlation values, NN, for each voxel is found and saved to construct the sparse adjacency matrix for the functional graph. In order to construct an undirected graph, the adjacency matrix is made symmetric, and hence the NN for each voxel will no longer be the same, and the number of nonzeros (nz) will increase.

Since we aim to compare the geometric and functional graph setup, the number of possible graph connections/number of nz should be approximately the same. From the number of neighbors in the geometric setup, the nz follows from Figure 4.6. By determining the nz in a functional setup using the same number of neighbors as in the geometric setup, a correspondence between the geometric and functional non-zeros can be found, as in Figure 4.7. Thereby the nz in the functional graph can be downgraded by use of a different number of neighbors; see the spy plot of the adjacency matrices of the functional graphs in Figure 4.8. In order to simplify the notation of which graph size is used, the geometric NN will be used in the following to define both the function and geometric graph size.

Tested Functional Graph Manipulations

Adjusting the Number of Nonzeros

Even with the correction to control the number of connections in the graph, the functional setup still fosters more connections, again following from the symmetric adjacency matrix. In order to achieve a more even number of connections per voxel, removing the weakest links is tested by checking the correlation values. The procedure is performed for each row in the adjacency matrix, but requires a large amount of data storage. However, there is a risk that clusters can be divided into sub-clusters, and even create more single-voxel clusters, which are not of interest.

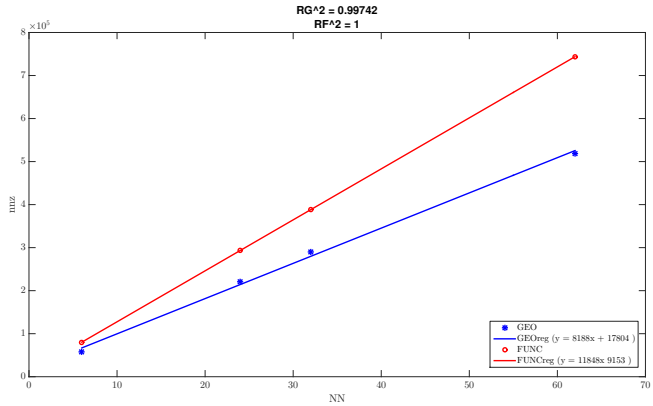


Figure 4.7: Determination of the number of neighbors of the functional graph based on the number of nz (on the simulated data). The regression has resulted in the following number of neighbors for the functional graph: 5, 17, 23, 44. Note that the gradient is not consistent among subjects, and must be computed for all datasets.

Cluster Density Filtering

As can be seen in Figure 4.5 (a.2), the functional resting state network is quite un-localized, and will thereby give rise to many different cluster connections, both when testing the correct paradigm, and during permutations. In order to promote a more stable cluster connection, we apply a geometrically based density filtering, see Figure 4.9. The correction is actually first performed after applying the t-threshold to achieve the cluster areas (Figure 4.5 (c.2)), and we seek to decouple functional clusters that are not strongly connected to a geometrically clustered area. The density count for a functional cluster is calculated from the geometric cluster areas, and only functional between-cluster links are included. No normalization regarding cluster size is performed, and a cut off is specified in order to filter out 'loosely' connected clusters. The cut off is half the number of neighbors (*i.e.* 3, 12, 16, 31), *i.e* two functional clusters will be decoupled, if there only exist NN/2 links between the geometrically specified areas. The filtering is performed for all permutations, and only on the extracted, activated areas. However, it is desirable to perform this non-traditional functional-geometrical combination on a global scale, and possible solutions will be discussed in Section 4.7.

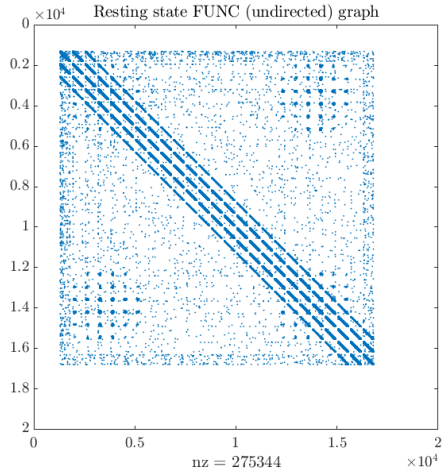


Figure 4.8: Example of a functional graph for the simulated data (and constructed to correspond to a geometric 32 neighborhood setup). Note the visible cluster connections in the corners, corresponding to the linkage of the diagonal clusters in Figure 4.4. There is also a clear connection to the neighboring voxels specified by the high density areas near the diagonal.

Using the Highest Percentage

Going away from the process of taking the row-wise NN highest values in the covariance, using the max percentages of the covariance has been considered. This would create an easy, controllable number of non-zeros (easy to match with the geometric graphs), but require a large amount of data storage and bookkeeping. Another pitfall is the likelihood of creating hubness in the data, without capturing the desired effect. The procedure will therefore not be used in the following.

(b) Statistic Under Activation

In Figure 4.5 (b) the t-statistic under the stage 2 data is performed, as described in Section 4.3.2. The figure holds the left hand stimuli depicted together with the design matrix, holding the corresponding HDR response for both stimuli situations. The contrast depicted below specifies the test performed, *i.e.* left hand movement against the rest. By use of the GLM, the t-heatmap in (b) is constructed. However, it is necessary to specify a number of voxels (a maximum

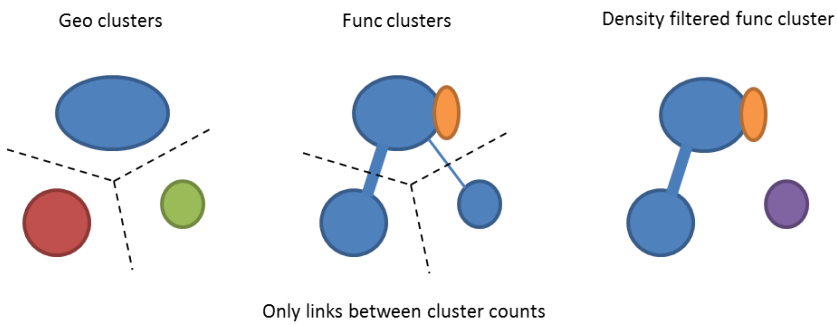


Figure 4.9: Proposed cluster-density filtering. The first sub-figure illustrates the geometrically defined clusters, with the dotted lines indicating separated areas. The second sub-figure describes the functional clusters, consisting of two clusters, a blue and an orange. By applying the geometric areas from the first plot, the link density between areas can be calculated. The link density is visualized by the thickness of the blue intersecting lines. The functional orange clusters that exist in the same geometric area as the blue have no link to the defined cluster areas, and will be left untouched. In the last sub-figure, the cluster with low link density (purple cluster) has been disconnected from the blue cluster, and only strong linkage between clusters exists.

t-value) that have to be investigated as clusters, [Nichols and Holmes, 2001]. The corresponding cut-off t-value is used in the following permutations. Use *e.g.* a group histogram over t-values in all subjects to specify an appropriate t-level/number of voxels to investigate.

(c) Graph Connections and Thresholding

The following is performed for both the functional and geometric graphs in the 'active' state (under stimuli). The thresholded statistical parametric maps from 4.5.(b) is seen in the top of (c.2). The adjacency matrices of the resting state graphs are also depicted in (c.1), as well as the thresholded voxels used to extract subgraphs with only voxels above the specified t-level, as seen in (c.3), for the functional and geometric setup. The corresponding 3D 'glass brains' can now be used to investigate the performed clustering, as in Section 4.3.4. The accumulated t-statistic is calculated for each cluster, as in (c.4).

In Figure 4.10, the subgraphs and the 3D glass brains can be seen for the simulated data. Figure 4.11 holds all the tested NN for both the functional and geometric setup. Note how the connected components grow when NN increases.

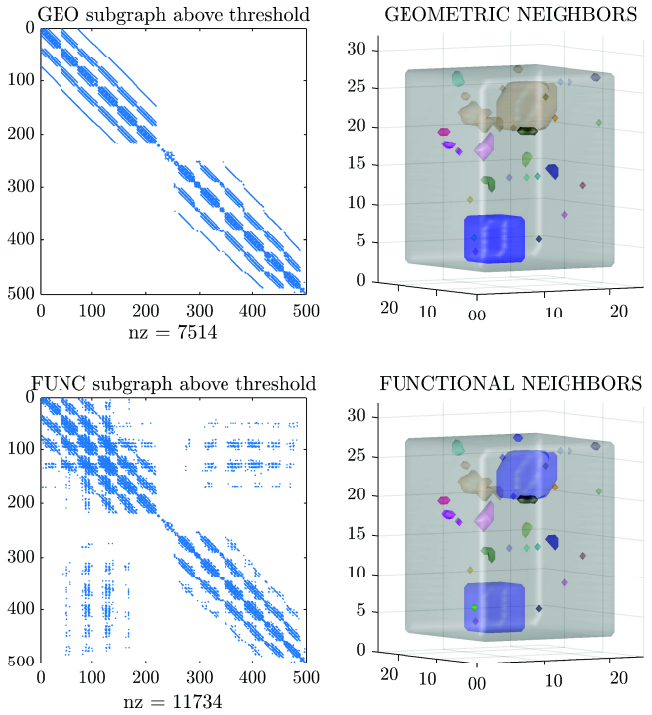


Figure 4.10: Geometric and functional subgraphs, and the corresponding connected clusters. The color-coding is used to specify connected clusters. Note that the two areas with co-varying stimuli are connected in the function graph setup. The setup with 24 NN has been used for all NN results, see Figure 4.11.

(d) Permutations

The goal is now to assess the significance of the previously identified cluster t-values. To do so, the permutation distribution is approximated by the permutation histogram in order to assess the p-values. By permuting the labels, here the stimuli, a new design matrix can be designed ((d).1) and new t-values are calculated ((d).2). By applying the global t-level ((d).3), the new subgraphs can be extracted, and the connected clusters can be found ((d).4). Since the theory behind permutation tests gives the freedom to choose any test statistic on hand,

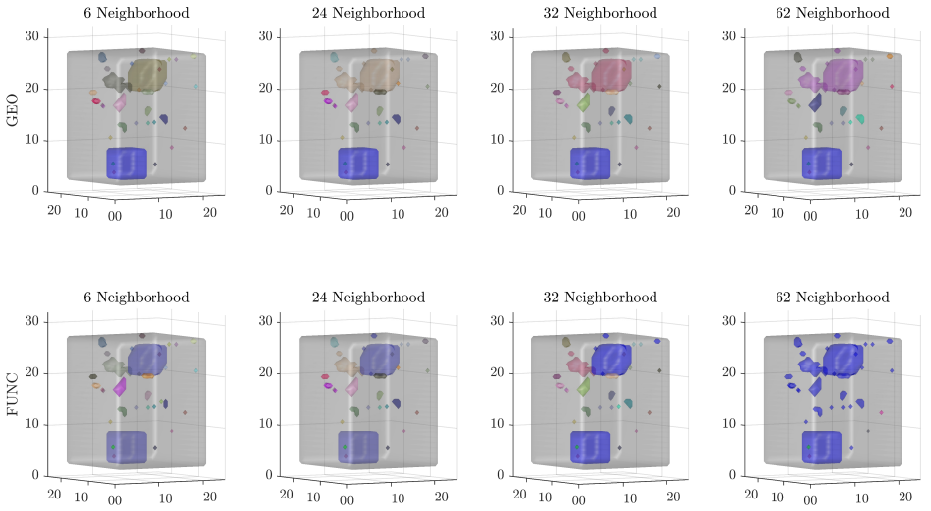


Figure 4.11: Geometric and functional 3D glass brains on all clusters above t-threshold. The clusters are color-coded according to the connected clusters. A higher number of neighbors features more connected clusters in both the geometric and the functional settings.

it is decided to use the largest aggregated t-value, as in [Nichols and Holmes, 2001]. For each permutation, the largest cluster's t-values give one datapoint for the histogram (see Figure 4.2), and after *e.g.* 1,000 label permutations, a reasonable histogram approximation can be found.

Quantitative Performance Measurement

In order to compare and evaluate the performance on different data, a quantitative performance measurement is defined. The performance measurement, named P-fraction (P) is given by:

$$P = \frac{t_{max} - t_\alpha}{t_\alpha}, \tag{4.9}$$

where t_{max} is the cumulative t-value in the largest cluster, and t_α is the t-values for a significance level of $\alpha = 5\%$ on the permutation histogram. Normally, when assessing the significance, the proportion of permutations above the observed t-value is used to reject the H_0 hypothesis if it is above the critical α level. Since it is not guaranteed to have any permutations above the observed t-value, it is proposed to specify a P-fraction in order to account for the size of the t-value,

and the distribution of the histogram. A high P-fraction is desirable, and will specify the low probability of observing permutations above the t-value. Please refer to Section 4.7 for more discussion on the use of P-fraction.

4.6 Results

4.6.1 Simulated Data

The resulting permutation histograms for the simulated data can be seen in Figure 4.12. Cluster t-values are indicated by letters, and clusters are significant with $\alpha = 0.05$ when they are above the red line indicating the t_α value. The reported p-value for the highest cluster is seen to be $< 1/1,000$, since no permutation yields a higher t statistic. Note that two clusters are significant for the geometric graphs setup, where only one larger significant cluster remains in the functional version. The difference in the values for increasing NN is more clearly illustrated in Figure 4.13, where the significant clusters are visualized in the 3D glass brain.

From Figure 4.13, it is apparent that the excepted active areas can be found regardless of the settings, but the geometric setup finds it as two disconnected clusters. It is clear to see how NN in the resting state graphs affects the connected cluster (even though there is not a high difference between 24 and 32). When increasing NN in the geometric setting, nearby clusters are being connected, resulting in a more diffuse cluster representation. The number of significant voxels is increasing, but the P-fractions are not considerably different NN in between. Note, however, that the P-fraction for the second cluster is much lower than that of the first. For the functional settings, there is only one significant cluster, since the two active areas are functionally connected by the graph. The increase in NN is observed to connect nearly all clusters above the t-level (see Figure 4.13 and Figure 4.11), inducing a large amount of false-positive errors.

The results reported above were obtained without density filtering, so in order to visualize the effect, Figures 4.14 to 4.16 hold the density filtered results. Comparing Figure 4.11 and 4.14, a clear difference is seen in the number of connected clusters for the functional clusters. The link density filtering has definitely decoupled multiple clusters - especially for a higher number of NN. Nearly no tail is seen on the permutation histograms in Figure 4.15, entailing a higher p-value and a much lower t_α levels, compared to Figure 4.12. Note that the significant cluster depicted in Figure 4.16 can therefore be described by a much higher P-fraction. For the setting with 24 NN, two significant clusters are now produced in the functional setup up as well, due to the decoupling of weakly connected clusters.

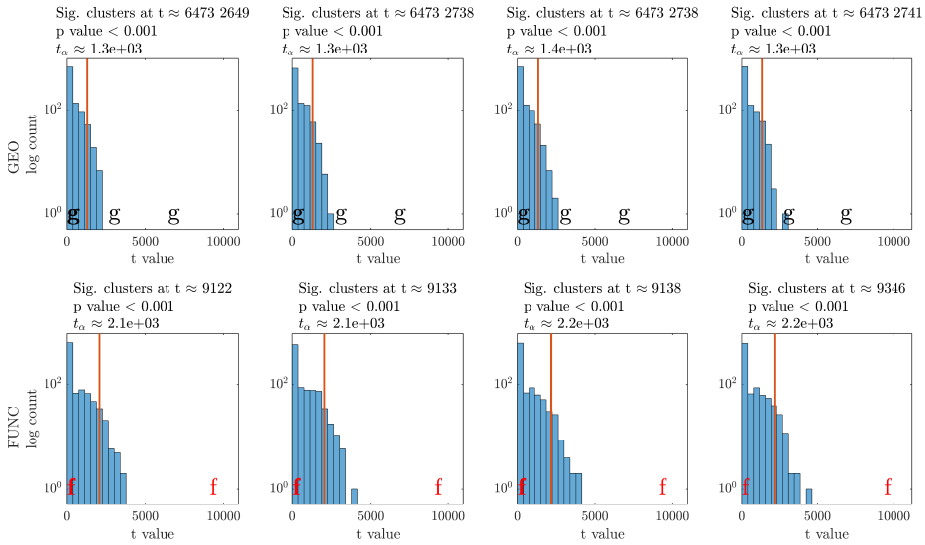


Figure 4.12: Permutation histograms over max t-value (simulated data) for both geometric (top row) and functional configurations (bottom row). Note the semi-logarithmic y-axis in order to display the tail of the histogram. The superimposed letters symbolize the t-value for the clusters to investigate (as seen in Figure 4.11). The red vertical line indicates the t_α level, and clusters above this line are significant, and depicted in Figure 4.13. The title of each subplot displays the t-values for (up to the three highest) significant clusters, together with the p-values of the largest cluster and the t_α value. The different columns correspond to the NN setting.

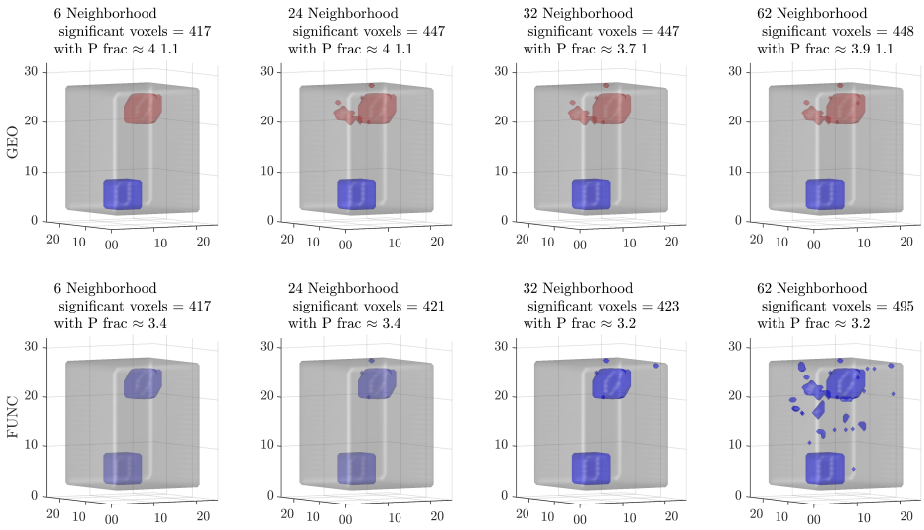


Figure 4.13: 3D glass brain of significant clusters in the simulated data. The subfigure setup in geometric/functional vs NN results is maintained, as in Figure 4.12. Color-coding indicates connected clusters (with no cluster size dependency), and the title specifies the number of significant voxels and the P-fractions for significant clusters.

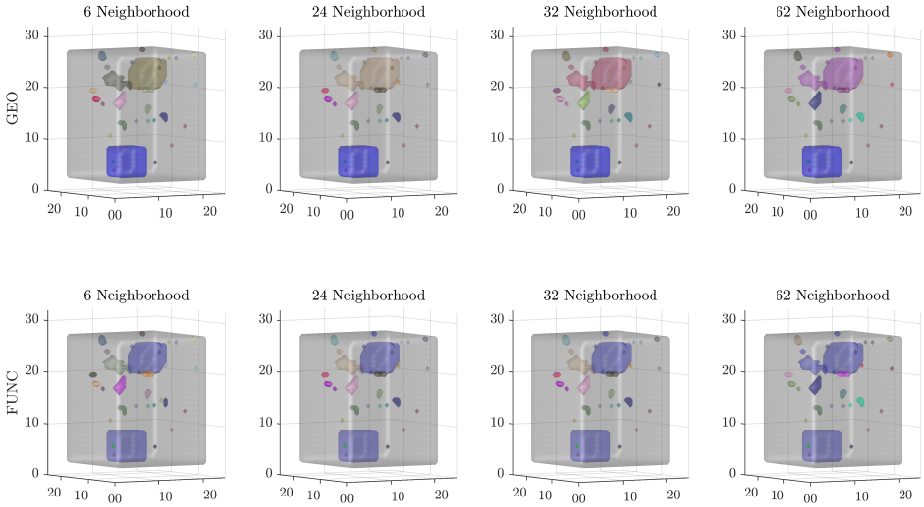


Figure 4.14: Geometric and functional 3D glass brains on all clusters above t-threshold, when applying link density filtering. Compare to Figure 4.11.

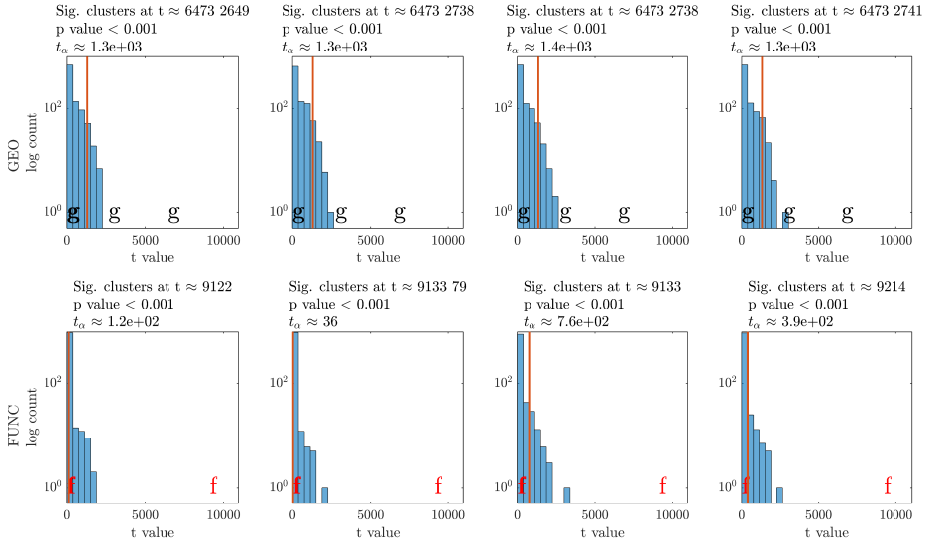


Figure 4.15: Permutation histograms (simulated data) for both geometric (top row) and functional configurations with link density filtering (bottom row). Note the semilogarithmic scale, and the titles t, p, and t_α values, and the setup as described in Figure 4.12.

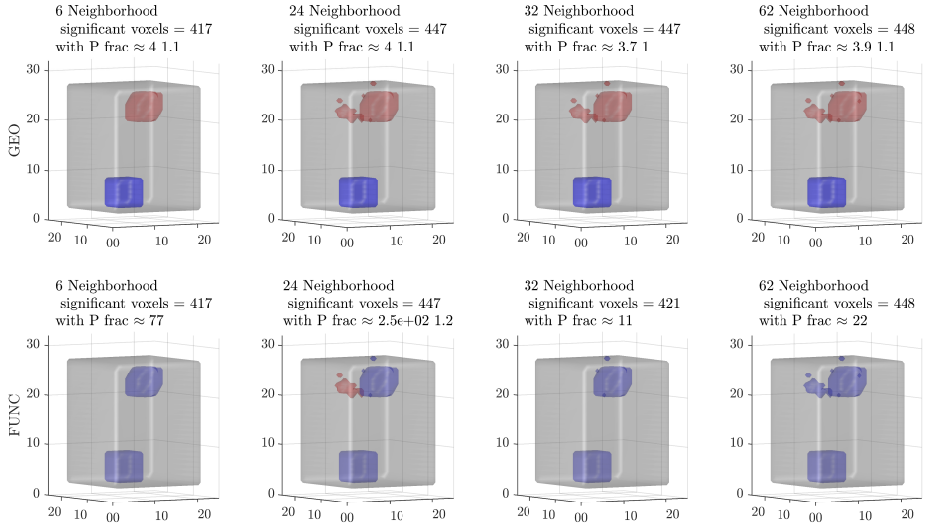


Figure 4.16: 3D glass brain of significant clusters in the simulated data with link density filtering. The figure setup remains the same as in Figure 4.13.

4.6.2 Finger Tapping Data

Figures 4.17 to 4.20 hold results from two subjects from the finger tapping dataset. 1,000 voxels were chosen to investigate and 1,000 permutations were performed. The same graph setup as for the simulated data is used, and link density filtering is performed, but the effect of applying the filtering was not as effective as for the simulated data. For both subjects, there is a clear activation in the motor cortex corresponding to the left hand movement, as well as a counter cerebellar activation. These areas are known to co-vary [Rasmussen et al., 2012], and are seen to be connected in the functional setup. It appears that the geometric constellation is less sensitive to the choice of NN than the functional. Again, the functional constellation tends to connect all clusters when using a 62 NN setup. It is also possible to see the diversity between subjects with respect to the amount and size of the extracted clusters.

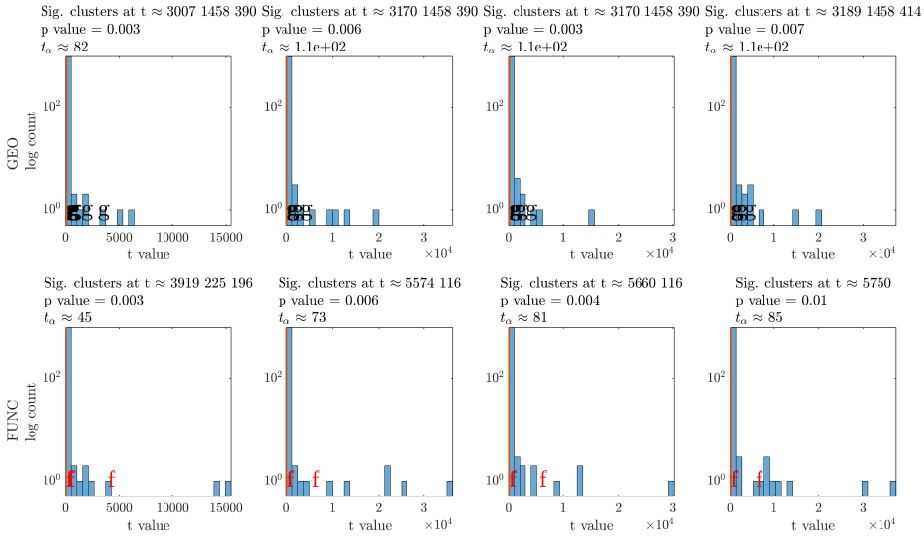


Figure 4.17: Permutation histogram for Subject A. Note the semilogarithmic scale, and the titles t, p, and t_α values, and the setup as described in Figure 4.12.

To summarize, all subjects' performances Figure 4.21 give a view into the performance of the functional setup compared with the geometric. The four bar charts represent the test results from different choices of voxel numbers to investigate (750, 1000, 1250 and 1500 voxels). The colors in the bar charts specify the distribution of the 28 subject results, and are described as follows: Dark blue corresponds to the functional graph setup having the largest P-fraction and the highest number of significant voxels. The light-blue specifies an equal

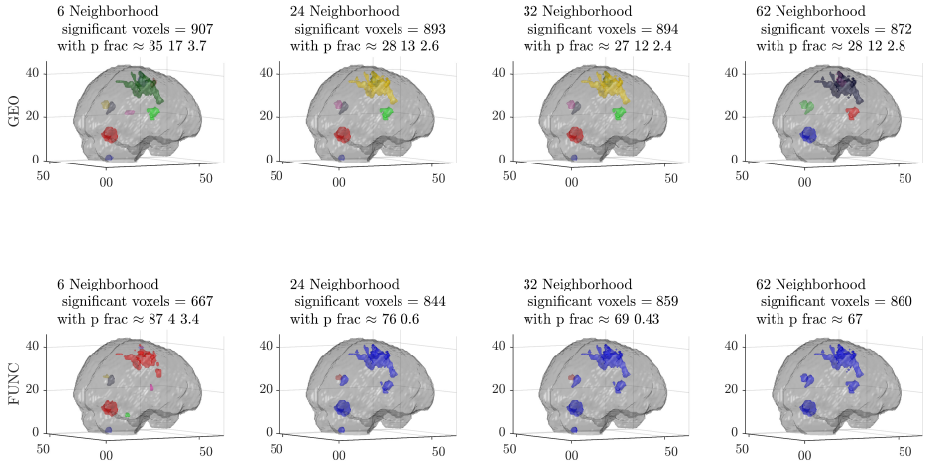


Figure 4.18: Significant clusters in Subject A. The figure setup remains the same. Note that the shift in cluster color is of no importance - the cluster color is only interpretable for one test.

performance of both the functional and geometric setup. The turquoise areas mark the number of subjects in which the geometric setup gives the highest number of significant clusters and the highest P-fraction. The green number of subjects are when the geometric setup has the highest number of significant voxels (GV), but the lowest P-fraction (FP). In contrast, the orange number is when the functional setup has the highest number of significant voxels (FV), but the geometric setup has the highest P-fraction (GP). A number of subjects are also marked as having no significant clusters at all, and these are increasing for higher NN.

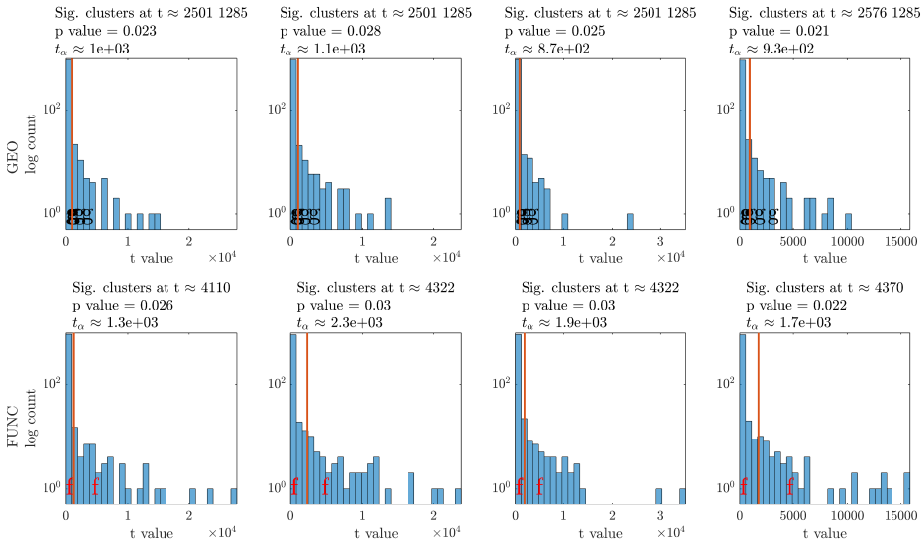


Figure 4.19: Permutation histogram for Subject B. Note the semilogarithmic scale, and the titles t, p, and t_{aplha} values.

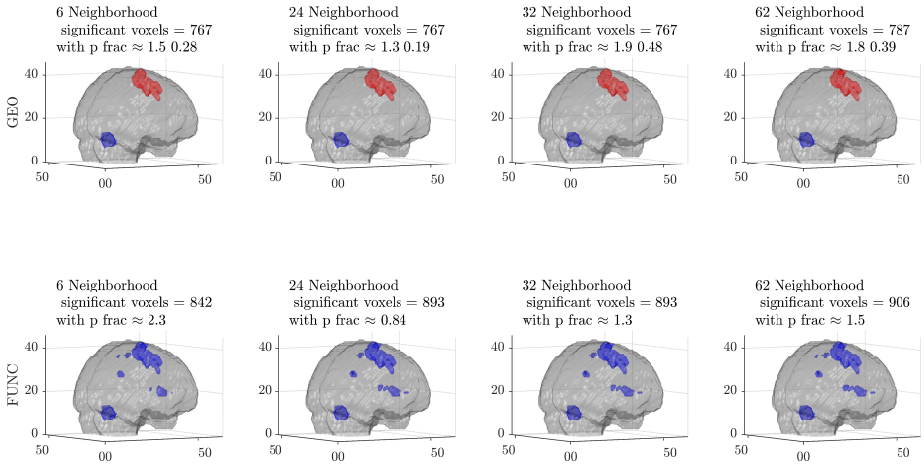


Figure 4.20: Significant clusters in Subject B. The figure setup remains the same. Note that the shift in cluster color is of no importance - the cluster color is interpretable for one test only.

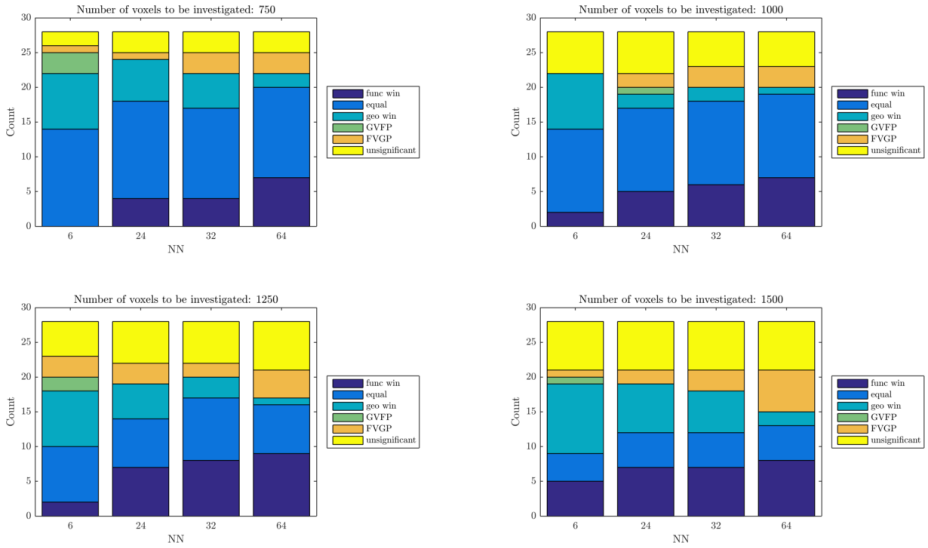


Figure 4.21: All subjects' performance evaluated based on P-fractions and number of significant voxels (as specified by the colors), for four different settings of the number of voxels to investigate, and for the four settings on NN. For in-depth description, please see the text.

4.7 Discussion

The preceding has summarized the current workflow of the proposed GBCPT, but there are still many aspects that need to be considered. The following will discuss some of the current results, but also touch upon future possibilities of development.

When reviewing the results of the simulated data, the hypothesis of connecting areas to gain a high test statistic is very clear. However, the functional permutation histogram will also tend to be connected under a permutation, and can therefore entail a higher permutation value than an equivalent geometric setup will foster. Link density filtration drastically improves the results for the simulated data, but for the real data, the reported results were not as evident. This could be due to the link cut-off being set too low for real data, and could be investigated further. The strong results from the link-density-filtered simulated data could be due to slightly too strong filtering, implying that nearly no clusters are functionally connected under permutations. This could also be due to the simple nature of the simulated resting state data having little connection except in the predefined areas. Other improvements could include a cluster size filtration and a further combination of a graph holding both functional and geometric connections.

The introduced P-fraction can take the difference in the permutation histogram into account, and present an arbitrary value that is more robust than the p-value *e.g.* if both the geometric and functional setups had a $p<0.001$ (no permutations over the t_{max}), the P-fraction can indicate which setup is most reliable. A high P-fraction indicates a low probability of observing permutations with values above the t_{max}. It should be noted, however, that the P-fraction is slightly in favor of the geometric setting, since t_α in the real data tends to be lower than the functional level. When comparing the setups, it is also of great importance to account for the fact that the geometric design gives rise to many separate significant clusters. Even though Nichols and Holmes [2001] uses the largest-cluster-test-statistic histogram produced to assess the p-values of the smaller cluster, it could be argued, that a new histogram based on second-largest-cluster-test-statistic must be produced in order to more correctly report the p-value of the second largest cluster (and so on). In this work, the largest-cluster-test-statistic histogram will still be used to assess the P-fractions of all the significant clusters. Another way to describe the geometric settings' multiple significant clusters is to consider taking the mean P-fraction over all the significant clusters - which could give a more fair comparison between the geometric and the functional setup. With the previous in mind, the simulated data shows a proof of concept of the GBCPT, both with and without link-density filtering, and gives rise to considerations about the number of neighbors chosen.

When considering the NN influence on the simulated data, Figure 4.13, the geometric setup already produces type 1 error areas when increasing to 24 NN. The 24 and 32 NN geometric setup are very similar to the spatial extent, see Figure 4.6, and will therefore often produce almost similar geometric results. In the functional setting, an NN of 64 nearly connects all clusters, resulting in many false positives. On that note, the problems with a fully connected graph should be discussed. The setup has been tested for both simulated and real data (not shown), and it naturally shows equal results for the geometric and functional settings, and a generally lower P-fraction. The reported number of significant voxels (as also reported in Nichols and Holmes [2001]) is therefore a number that should be seen in connection with *e.g.* the P-fraction, since a fully connected graph will obviously create a high number of significant voxels.

It should be noted that it is not possible to translate the simulated experiments directly to the real data. Even though the simulated data is designed to mimic simple fMRI behavior, limitless differences exist. Still, more simulated experiments should be performed, as they facilitate ground truth results for a frame of reference. More tests could be explored by investigating a specific number of voxels, as done for real fMRI data. It would also be obvious to report the false positives and false negatives to estimate performance.

Another aspect is the size of the cluster areas sought for, which are resolution dependent. If the goal is to estimate the significance for a *e.g.* 500 voxel area, the allowed number of neighbors must be set appropriately (compared to a 50 voxel area). A too low number is not likely to connect two distant clusters, since the NN can only cover inter-cluster connections.

The difference between the areas of the two subjects shown is most likely due to their different t-statistics. The chosen number of voxels is the cause, as different numbers of voxels will create very different cluster areas to investigate. In an SPM plugin (under construction), it would be convenient to use the directly calculated SPM t-statistic (includes a more comprehensive design matrix), and investigate other methods to extract a correct threshold.

When considering the result in Figure 4.21, all the above comments on P factions and number of significant voxels should be taken into account. It is therefore hard to directly define what functional or the geometric setup is to be preferred over another. In fact, the work done should not be seen as a comparison between the geometric and the functional constellations, but more as proof of concept of a graph-based cluster permutation test. Permutations in general have the drawback of the processing time to construct the histogram, and the GPICA has an extra time dimension due to the creation of the resting state graphs. Our proposed procedure of permutation of the stimuli is questioned by Nichols and Holmes [2001], but we claim that exchangeability is still fulfilled under the H_0

hypothesis.

The overall aim is to develop a plugin for SPM that offers the possibility of a graph-based cluster permutations test based on resting state connectivity graphs. Collaborative work could be done to supply a standard graph for resting state data. Either produced as a 'simple' mean over all subjects, or with incorporation of brain connectivity over different regions. An article based on this chapter is therefore planned to be completed. Additional applications of GPCPT can also be investigated, since the graph based setup are not only limited to fMRI, and a possible utilization could be AB testing with a graph based prior from social networks.

Functional Graph Improvements

The issue regarding the functional graph adjustments should perhaps be dealt with in an entirely different way (since the link density filtering procedure can be questioned as the adaption of the graphs is only done on the subgraphs). Our proposed solution is constructed to give each voxel (nearly) the same amount of connections, but work on whole-brain connectivity could be the background for changing the functional graph development. By defining more grouped areas, *e.g.* through ICA, [Damoiseaux et al., 2006], or by Infinite Relational Modeling, [Møup et al., 2010], to extract coherent groups in the resting state networks, better functioning resting state graphs for permutation could be developed. Extending the coherence to be based upon Mutual information could also serve as an improvement, and more brain connectivity studies can be investigated.

4.8 Conclusion

A graph based cluster permutation test has been proposed for fMRI data, with the clustering information extracted from resting state networks. The concept was proven by achieving a sensitivity for the highest cluster statistic. The great room for improvement in specifying the the function graphs, opens for achieving even more prominent results. The high sensitivity that can be attained is desirable for neuroscientists when testing for various conditions to improve our understanding of the brain. By incorporation prior information in graphs, cluster permutation tests could now be applied on a wide range of fields due to the diversity of the graph applications.

Chapter 5

Discussion and Conclusion

The main topic of this thesis has been two fold, focusing on a unsupervised and a supervised method to improve the analysis of fMRI data. First, an unsupervised ICA incorporating temporal source prior is developed (GPICA). The temporal dependency is facilitated by incorporation of Gaussian process source priors, and feature extraction of smooth reliable sources. Concurrently representing known noise patterns of the data. In the second part, covering supervised improvement, the application of cluster permutation tests is expanded. By extending the model to be graph based (GBCPT), multiple cluster possibilities arise. Through incorporation of functional brain connectivity, in the network, it is possible to improve the test statistic, when testing for functional significant areas in the brain. Work is still conducted to increase the significance of the improved sensitivity, but appears more than promising.

Incorporating prior knowledge on neural responses in GPICA, would be appropriate for EEG applications, where an even higher temporal dependency is present, due to the higher sampling rate. A toolbox for GPICA on EEG is therefore under development, whereto high performance expectations exist. A better, prior based ICA on neuroscience data will help improving our knowledge of the brain processes. The convolutive extension to GPICA is a further expansion, providing an extra dimension in which the propagation of the brain processes can be investigated.

Considering the GBCPT, promising results have been achieved, by combining well-known cluster permutation test with state of the art methods of extracting clusters in resting state networks, of the human brain. By improving the quality of the statistical analysis, identified contributors will be of greater statistical significance and thus provide valuable insight that would otherwise be considered to be nothing more than a trend. That could in the end improve our knowledge of the human brain and *e.g.* neurological diseases.

To tie together the two methods proposed, it should be investigated whether GPICA extracted spatial maps on resting state data, can be used to specify clusters in the resting state network for the GPCPT. As the GPICA is developed to support the nature of the BOLD signal, it would indicate that the inferred components capture some of the temporal dynamics in coherent brain regions under resting state. Improvements on both methods are still being carried out.

This thesis also proves, that the fields of ICA on fMRI data and statistical testing are far from exhausted, and that ICA is an informative tool even on single subject data. However, both proposed analysis methods are relatively slow with a high memory requirement, but thanks to the drastic rise in computational power and data storage, it is not considered to be an issue compared to the gain in knowledge and performance.

Returning to the opening line: "All models are wrong, but some are useful" [Box et al., 1987]; Stating that all models have various model assumptions, and in even the most simple cases, assumptions that can be hard to fulfill. In general it is hard work getting models to work, considering both model assumptions, model parameters and multiple model evaluation possibilities. The search for the optimal data, the optimal model, and the optimal solution will therefore never end; but keeping these limitations in mind, the development in science and machine learning can take us in the right direction.

Appendix A

Gaussian Process Based Independent Analysis for Temporal Source Separation in fMRI

Ditte H Hald, Ricardo Henao, and Ole Winther. Gaussian Process Based Independent Analysis for Temporal Source Separation in fMRI. *Submitted to NeuroImage*, 2016.

Gaussian Process Based Independent Analysis for Temporal Source Separation in fMRI

Ditte Høvenhoff Hald[1], Ricardo Henao[1,2], Ole Winther[1]

1) DTU Compute B321
Technical University of Denmark
DK-2800 Lyngby, Denmark
2) Institute for Genome Sciences & Policy
Duke University
Durham, NC 27708, USA

Abstract

Functional Magnetic Resonance Imaging (fMRI) gives us a unique insight into the processes of the brain, and opens up for analyzing the functional activation patterns of the underlying sources. Task inferred supervised learning with restrictive assumptions in the regression set-up, restrict the exploratory nature of the analysis. Fully unsupervised independent component analysis (ICA) algorithms, on the other hand, can struggle to detect clear classifiable components on single subject data. We attribute this shortcoming to inadequate modeling of the fMRI source signals, by incorporating a temporal source prior. fMRI source signals, biological stimuli and non-stimuli related artifacts, are all smooth over a time-scale compatible with the sampling time (TR), and we therefore propose Gaussian process ICA (GPICA), which facilitates temporal dependency of the extracted sources, by use of Gaussian Process priors.

On two fMRI data sets with different sampling frequency, we show that the GPICA inferred temporal components, and associated spatial maps, that allow for a more definite interpretation than standard ICA methods. The temporal structure of the sources are controlled by the covariance of the Gaussian Process, specified by a kernel function with a interpretable and controllable temporal length scale parameter. We propose a hierarchical model specification, considering both instantaneous and convolutive mixing, and infer

source spatial maps, temporal patterns and temporal length scale parameters by Markov Chain Monte Carlo. A companion implementation made as a plug-in for SPM can be downloaded from https://github.com/dittehald/GPICA.

Keywords: Gaussian Processes, fMRI, Source Separation, Independent Component Analysis, Convolutive Mixing, Bayesian Inference

1. Introduction

The hemodynamic response (HDR) of the brain, and the emerged blood-oxygen-level dependent (BOLD) contrast image captured by functional magnetic resonance imaging (fMRI), is a window into many aspect of the brain. In order to analyze fMRI signals, independent component analysis (ICA) is one of the most common ways to extract signal constituents. These sub-parts of the fMRI signal arise from different independent processes related to the stimuli, non-stimuli effects (such as heart beat) and artifacts (such as head movement), (Duann et al., 2002, McKeown et al., 2003). Proper source separation will thus allow both identification of stimuli related signals and artifact removal. These independent processes are expected to vary smoothly over time, on a time scale that is often comparable with the fMRI acquisition frequency. It therefore seems obvious to use this information in the modeling of the sources. This is, however, computationally involved, and ICA models based on (non-Gaussian) independent identically distributed (iid) sources are the most common used in practice, *e.g* Infomax and FastICA (Bell and Sejnowski, 1995, Hyvärinen, 1999).

The fMRI data, X, is assumed to be a linear combination of stimuli, non-stimuli and artifact signals plus additive noise. Each signal can be thought as consisting of a source, Z, and its dispersion in space, W, by $X = WZ + \epsilon$, with ϵ being the noise contribution. The estimation of W and Z is identifiable (up to permutation and sign symmetries) for some choices of prior distributions on the model parameters (Kagan et al., 1973, Henao and Winther, 2011). The non-identifiable case is an i.i.d. Gaussian priors on Z and W corresponding to probabilistic PCA and standard factor analysis. In blind source separation, ICA is the most widely used generative model to solve the problem, and most notably

identifiable model specification is based upon non-Gaussian i.i.d. priors or temporally correlated Gaussian priors. Most common used algorithms using the non-Gaussian i.i.d. priors either explicitly, or implicitly, are inter alia InfoMax (Bell and Sejnowski, 1995) and FastICA (Hyvärinen, 1999). The Molgedey-Schuster algorithm (icaMS) (Molgedey and Schuster, 1994) is a notable example of a model that implicitly uses a temporally correlated source prior.

In this paper we work with temporally smooth sources. We accomplish this by modeling each source as a temporal Gaussian process. In Gauss processes (GP) (Rasmussen and Williams, 2006), the covariance functions shape and parameters, determine the length-scales on which the signal is correlated. Consequently, in our algorithm, Gaussian process ICA (GPICA), different sources will differ both in terms of their spatial and temporal patterns, as in standard ICA, and additionally by the length-scale of their characteristic temporal correlation. As a part of the algorithm we infer W and Z and for each source the length-scale parameter of its kernel function.

A prerequisite for this model to make sense is that the sampling time (TR, the inverse of the sample frequency) is on the same length-scale as temporal correlation in the signal components we would like to recover. Typically the fMRI acquisition frequency is around 0.5-1 Hz (that is TR $\approx$ 1-2 sec.). We consider two data sets with different sampling rates: a fast acquisition visual paradigm dataset with TR = 1/3 sec. and motor paradigm dataset with a more standard TR = 2.49 sec. In both cases, we show that GPICA can recover clearly interpretable time scale signals, compatible with the TR used.

The Gaussian process has the property that it reverts back to independent Gaussian variables (that is probabilistic PCA), if the temporal length-scale parameter in the covariance function is much shorter than TR. We can in principle still get identifiable source recovery in this limit as long as the prior over the W matrix has a non-Gaussian distribution. Therefore we use i.i.d. Student's t-distributions on the elements of W. This choice has the added benefit of promoting more discriminative W because relative to the Gaussian, Student's t has more shrinkage towards zero of non-important parameters and less for important ones.

Gaussian processes (GP) are widely applied in inter alia regression problems, and has previously been employed into factor analysis (FA), where Yu et al. (2009) uses time point FA tied together by a GP in order to perform dimension reduction and smoothing simultaneously. Luttinen and Ilin (2009) developed a GPFA algorithm for reconstructing missing data, with GP attached to both factors and loadings. They uses a variational Bayesian framework for learning the model, and factorize the posterior approximation either in time or space - due to model complexity. (Luttinen and Ilin, 2009) are closely related to the GP work by Schmidt et al. (Schmidt and Laurberg, 2008, Schmidt, 2009). Schmidt and Laurberg (2008) focuses on probabilistic non-negative matrix factorization (NMF), and models factor smoothness with a GP. In the follow-up work, Schmidt (2009) uses non-linear mapped GP's (Warped GP's, (Snelson et al., 2004)) to perform function factorization. Bayesian inference is performed with Hamiltonian Markov chain Monte Carlo. Preliminary work on source separation with GP sources has also been formulated by Park and Choi (2007, 2008). Park and Choi (2007, 2008) use mutual information minimization (Park and Choi, 2007) or Gradient-based optimization (Park and Choi, 2008) of the log-pseudo-likelihood to infer the mixing matrix. Their proposed model is closely related to our GPICA model. Our novel contributions are the hierarchical specification of the model, Markov Chain Monte Carlo (MCMC) based inference, application to fMRI datasets and a companion plug-in for SPM to make the algorithm available for other researchers. The algorithmic workflow is illustrated in Figure 1. We also propose a convolutive mixing matrix extension of the algorithm. This extension is related to the work by Olsson and Hansen (2006). The main difference in the model specification is that the temporal model is a linear state space model (AR model), that may be seen as a different parameterization of the GP temporal covariance (Hartikainen et al., 2010). Furthermore, the model was not hierarchical and inference was carried out with expectation-maximization (EM).

The proposed algorithm has relatively high complexity, mainly due to the number of MCMC samples needed to get sufficiently accurate estimates. However, our empirical results show that this is sometimes a price worth paying. For minimally pre-processed data we can recover stimuli related sources even in cases where the amplitude of these sources

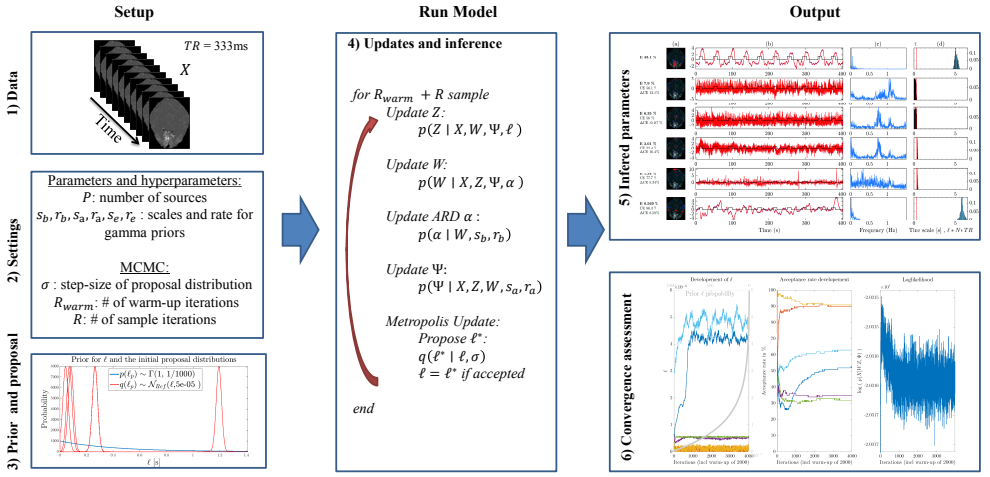


Figure 1: Schematic figure of GPICA, including model setup, concatenated updates and inference formulas, and the common output for single-slice data. Setup includes 1) data and knowledge of TR, 2) Specifying parameters and MCMC settings. In 3) the prior for ℓ is visualized together with the initial proposed reflective Gaussian distributions with the specifies step-size. 4) symbolize running the model with updates of the inference scheme and the Metropolis sampling. The model output covers 5) the inferred parameters, and 6) the convergence assessment. For the sake of simplicity, single-slice data is used for visualization in this schematic form.

are only a small fraction of the total variance in the data. Standard ICA algorithms applied to the same datasets are not showing the same ability to recover meaningful sources. The careful time-consuming modeling and inference in GPICA are, without a doubt, contributing factors to these results.

The remainder of the paper is organized as follows: In Section 2, we introduce the GP-based ICA model with GP sources and both instantaneous and convolutive mixing. Appendix AppendixA gives the details of the Gibbs sampling based inference scheme. In Section 3 we give a brief description of the two fMRI data sets used in Section 4. In Section 4.5 we describe an implementation, as a plug-in to SPM. In Section 5 we interpret the empirical results. In Section 6 we provide an outlook and perspective for the model.

2. Independent Component Analysis with Gaussian Process priors

In this Section we will describe the independent component analysis (ICA) model with Gaussian process based sources. We will specify hierarchical Bayesian priors for model parameters and describe a Markov chain Monte Carlo (MCMC) framework for inference. We consider both instantaneous and convolutive versions of the mixing model.

2.1. Instantaneous Independent Component Analysis

The basic ICA model is defined as

$$X_{ij} = \sum_{p=1}^{P} w_{ip} z_{pj} + \epsilon_{ij} \tag{1}$$

$$X = WZ + \epsilon \ , \tag{2}$$

where X_{ij} is the data matrix with subscripts representing voxel and time sample, respectively. Each row in Z represents the time course of one source and W is the component map, or mixing matrix. P is the number of independent components in which the data is decomposed. P is much smaller than the number of voxels I and the number of time points J. ϵ_{ij}, the spatial-temporal noise contribution, is assumed to be iid Gaussian with zero mean and spatially varying noise variance Ψ_{ii}. We can thus write the likelihood for the jth timeslice as

$$P(x_j \mid W, z_j, \Psi) = \mathcal{N}(x_j \mid W z_j, \Psi) \tag{3}$$

and the joint distribution of the whole data set is

$$P(X|W, Z, \Psi) = \prod_{j=1}^{N} P(x_j \mid W, z_j, \Psi) \ , \tag{4}$$

where Ψ is diagonal with elements Ψ_{ii}.

2.2. Bayesian Prior Specification

We adopt a hierarchical Bayesian framework specifying priors on Z, W and Ψ. As in Park and Choi (2007, 2008) we use independent Gaussian process (GP) priors for each of

the P temporal source sequences $z_{[p]} = (z_{p1}, \ldots, z_{pN})^T$:

$$p\left(z_{[p]}|\ell_p\right) = \mathcal{N}(z_{[p]}|0, K_{\ell_p}) \tag{5}$$

$$p(Z|\ell) = \prod_p p(z_{[p]}|\ell_p) \tag{6}$$

where ℓ_p is shorthand for the hyperparameters of the pth GP. Formally, the GP specifies a joint Gaussian distribution for any finite collection of input variables $t_1, \ldots, t_J$. The covariance matrix is calculated using the covariance function $k_\ell(\cdot, \cdot)$. In this work, we will use a temporal covariance function: $K_{jj',\ell} = k_\ell(t_j, t_{j'})$, where t_j is the time of sample j with a squared exponential covariance function, with ℓ being a scalar temporal length-scale parameter:

$$k_\ell(t, t') = e^{-\frac{|t-t'|^2}{2\ell^2}} \ . \tag{7}$$

We will make a hierarchical model specifying a prior on ℓ_p. An ℓ_p much smaller than $\Delta t = t_j - t_{j-1}$ corresponds to an iid Gaussian prior, as the covariance matrix degenerates to $K_{jj',\ell} = k_\ell(t_j, t_{j'}) = \delta_{jj'}$. The model will thus fall back to factor analysis in that limit. A high ℓ induces a wider distribution around the diagonal, entailing more dependency on surrounding samples, and thus a function smoother on a longer time-scale. A high dependency of the previous samples will moreover not allow for high frequency content. Thereby ℓ can be considered as a frequency controlling parameter. Since ℓ is non-negative, the Gamma distribution is a convenient choice as prior for ℓ_p:

$$p(\ell_p) = \mathrm{Ga}\left(\ell_p|s_e, r_e\right) \tag{8}$$

$$p(\ell) = \prod_p p(\ell_p) \ , \tag{9}$$

where s_e and r_e are the scale and rate, respectively. Rough knowledge of the length-scale of the different processes (paradigm, blood flow due to pulse, movement artifacts, etc.) allow us to specify these parameters as discussed in Section 3 on the data sets used.

For the remaining parameters standard conjugate hierarchical priors with natural parameters are used. This means we work with priors over precisions = inverse variances.

The inverse noise variance, Ψ^{-1}, that is the noise precision, has a gamma distribution with a scale and rate adapted to match realistic biological signals (see data Section 3). This in turn corresponds to an inverse Gamma distribution for the noise variances (with rate inverted compared to the precision prior):

$$p\left(\Psi^{-1}\right) = \prod_{i=1}^{D} \mathrm{Ga}\left(\Psi_{ii}^{-1} \mid s_a, r_a\right) \tag{10}$$

$$p\left(\Psi\right) = \prod_{i=1}^{D} \mathrm{Ga}^{-1}\left(\Psi_{ii} \mid s_a, 1/r_a\right) \ . \tag{11}$$

Again s_a and r_a are the scale and rate, respectively.

For the mixing matrix, W, we would like to have a model that encourages more discriminative inferences, that is the elements should be allowed to become large, if necessary, but we also expect many to be close to zero. This can be achieved by a Student's t-distribution which has more probability mass in the tails and close to zero, than the Gaussian distribution. We use a hierarchical construction convenient for Gibbs sampling based inference to get Student's t-distribution elements of W: Elements are iid Gaussian with individual precisions α_{ij}:

$$p\left(W|\alpha\right) = \prod_{i=1}^{D}\prod_{p=1}^{P} \mathcal{N}\left(w_{ip} \mid 0, \alpha_{ip}^{-1}\right) \ . \tag{12}$$

The precision elements in turn has a gamma prior

$$p\left(\alpha\right) = \prod_{i=1}^{D}\prod_{p=1}^{P} \mathrm{Ga}\left(\alpha_{ip} \mid s_b, r_b\right) \ . \tag{13}$$

To see that this is equivalent to Student's t one may marginalize out the precision parameters:

$$p(w_{ip}|s_b, r_b) = \int \mathcal{N}\left(w_{ip} \mid 0, \alpha_{ip}^{-1}\right) \mathrm{Ga}\left(\alpha_{ip} \mid s_b, r_b\right) d\alpha_{ip} = t\left(w_{ip}|0, \frac{r_b}{s_b}, 2s_b\right) \ , \tag{14}$$

where $t(w|\mu, \sigma^2, \nu)$ is a Student's t-distribution with mean μ, scale σ and ν degrees of freedom.

This concludes the model specification. The graphical model representation is shown in Figure 2. To sum up, we can write the joint distribution of data X and model parameters

$$p(X, W, Z, \Psi, \ell, \alpha) = p(X|W, Z, \Psi)p(Z|\ell)p(\ell)p(W|\alpha)p(\alpha)p(\Psi) \tag{15}$$

as specified by Equations (4), (6), (9), (11), (12) and (13). The hyperparameters of the model are scale and rate parameters of the priors for GP covariance length-scale, the noise and mixing matrix precision. Posterior inference that is estimating $p(W, Z, \Psi, \ell, \alpha|X)$ is handled with MCMC as described in the next section.

2.3. Inference

Posterior inference can be handled with Gibbs sampling (iterative sampling of conditionals) for all parameters apart from ℓ because conjugate priors have been used. These are standard updates listed in AppendixA. Since the conditional distribution for ℓ_p, which is proportional to $p(z_{[p]}|\ell_p)p(\ell_p)$, Equations (5) and (8) is not in a standard family we instead adopt a Metropolis-Hastings approach: Propose new ℓ_p^* from proposal density $q(\ell_p^*|\ell_p)$ and accept the proposed value as the next value in the chain with probability:

$$\min\left(1, \frac{p(z_{[p]}|\ell_p^*)p(\ell_p^*)q\left(\ell_p \mid \ell_p^*\right)}{p(z_{[p]}|\ell_p)p(\ell_p)q\left(\ell_p^* \mid \ell_p\right)}\right) . \tag{16}$$

If not accepted the next value in the chain is the set equal to the current value ℓ_p. This step, that can be performed for all sources in parallel, requiring recomputing and inversion of the covariance matrix and is thus a $\mathcal{O}(P^3)$ operation.

The proposal distribution for ℓ_p, is chosen to be a so-called reflective Gaussian distribution i.e. the absolute value of a Gaussian variable with mean ℓ_p and variance σ^2:

$$q\left(\hat{\ell}_p \mid \ell_p\right) = \mathcal{N}\left(\hat{\ell}_p \mid \ell_p, \sigma^2\right) \tag{17}$$

$$q\left(\ell_p^* \mid \hat{\ell}_p\right) = \delta\left(\ell_p^* - \left|\hat{\ell}_p\right|\right) . \tag{18}$$

The reflective Gaussian is chosen to achieve the properties of the Gaussian random walk when ℓ_p is a few σs above zero while, preserving the restriction to only non-negative values.

We can show that the proposal is symmetric and thus reduces to the Metropolis algorithm[1]

$$\min\left(1, \frac{p(z_{[p]}|\ell_p^*)p(\ell_p^*)}{p(z_{[p]}|\ell_p)p(\ell_p)}\right) \ . \tag{19}$$

The step-size, σ, controls the properties of the Metropolis-algorithm and a step-size corresponding to an acceptance ratio around 10-20% is usually a good choice to aim for. We assess convergence diagnostics in Section 4.2.

2.4. GPICA Algoritmic Recipe

The resulting algorithm performs generative tracking of the different sources, giving detection of underlying independent components. The computational complexity is $\mathcal{O}(PJ^3)$ plus $\mathcal{O}(IJP)$, as the inverses of the covariance matrix and WZ has to be computed for each source. The algorithm, *Gaussian process independent component analysis (GPICA)* is freely available as a plug-in to SPM[2]. The MCMC updates for the model is given in AppendixA. The work-flow of the algorithm is schematized in Figure 1.

2.5. Summarizing Algorithm Output

After a preselected number of warm-ups, R_{warm}, (also known as burn-in) steps we take R MCMC steps and collect MCMC samples for all parameters. We index these samples by superscript (r) with $r = 1, \ldots, R$ and use medians to get point estimates, such as:

$$W^{\text{est}} = \text{Median}(W^{(r)}) \ ,$$

where Median$(\ldots)$ is element-wise median. We also report element-wise 5% and 95% quantiles when handling the point estimates of Z.

[1] We may write the proposal as a mixture of two truncated Gaussians: $q\left(\ell_p^* \mid \ell_p\right) = \left[\mathcal{N}\left(\ell_p^* \mid \ell_p, \sigma^2\right) + \mathcal{N}\left(\ell_p^* \mid -\ell_p, \sigma^2\right)\right]\Theta(\ell_p)$, where $\Theta(\cdot)$ is the Heaviside step-function. The first component corresponds to the non-reflected transition probability and the second term to the reflected. Both terms are symmetric to the exchange of ℓ_p with ℓ_p^*, and it is implicit that also ℓ_p^* is non-negative i.e. the proposal is symmetric.

[2] from https://github.com/dittehald/GPICA.git

In order to get an assessment of the relative contribution of the individual sources, we list the sources according to relative "energy":

$$E(p) = 1 - \frac{1}{||X||^2}\frac{1}{R}\sum_r \sum_{i,j}\left(X_{ij} - W_{ip}^{(r)} Z_{pj}^{(r)}\right)^2 \tag{20}$$

where the mean is subtracted from X and $||A||^2 \equiv \sum_{i,j} A_{ij}^2$. Since the model is additive, the sources will be complementary and the ordering and the relative energy should only taken as a rough indicator of importance. The cumulative relative energy (assuming that the sources are already sorted according to energy):

$$C(p) = 1 - \frac{1}{||X||^2}\frac{1}{R}\sum_r \sum_{i,j}\left(X_{ij} - \sum_{q=1}^{p} W_{iq}^{(r)} Z_{qj}^{(r)}\right)^2 \tag{21}$$

is also reported.[3] It should give a more reasonable assessment of the collective effect of the sources considered up to the current.

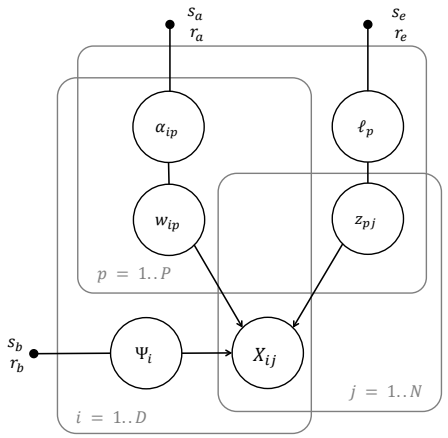


Figure 2: Graphical model for the GPICA.

[3] $C(p)$ is expensive to compute so we use $\widehat{C}(p) \equiv 1 - \frac{1}{||X||^2}\sum_r \sum_{i,j}\left(X_{ij} - \frac{1}{R}\sum_{q=1}^{p} W_{iq}^{(r)} Z_{qj}^{(r)}\right)^2$ instead. In practice we observe little difference between the two expressions.

2.6. Convolutive ICA

Convolutive approaches in ICA have been widely proposed in order to capture a potential complex temporal expansion of the sources (Lee et al., 1997, Parra et al., 1998, Olsson and Hansen, 2006). It has former been applied to especially EEG procedures and has shown to give a more comprehensive representation of the signals (Dyrholm and Hansen, 2004, Dyrholm et al., 2007). Convolutive ICA for fMRI has previously been investigated in inter alia Hansen (2003). It is therefore of interest to investigate if similar results can be obtained by a convolutive GPICA algorithm applied on simple fMRI data.

In convolutive ICA, data is assumed to consist of a convolutive mixture of the source signals (with added iid noise) with a mixing matrix W_τ for each time-lag $\tau = 0, \ldots, T$, where T is the maximal time-lag. The convolutive part is in this work performed purely in the time domain, and hence creates T+1 mixing matrices. Let x_j, z_j and ϵ_j denote the jth time slice of the data, sources and noise, respectively. Then the convolutive linear mixture model reads

$$x_j = \sum_{\tau=0}^{T} W_\tau z_{j-\tau} + \epsilon_j \ . \tag{22}$$

The MCMC algorithm for the convolutive model is implemented into the existing model by re-deriving the posteriors for W and Z, where Equation 3 is updated to encompass Equation 22. Apart from this extension, all model specifications remain the same. We will omit the derivation of the inference algorithm for brevity.

3. Data Description

To validate the proposed method we test the algorithm on two fMRI datasets; one single slice fast acquired dataset, and a full 3D volume with a longer repetition time, TR. We will restrict us to single subject fMRI in order to show the strength of the GPICA on single subject, single train data. We will test if the different paradigms from the two datasets (visual and motor skills, respectively) enables different evaluation of the extracted spatial maps.

3.1. Fast acquisition visual paradigm

The data is collected at Hvidovre Hospital (Petersen et al., 2000), from an experiment where a test person was subjected to a visual stimuli paradigm (flickering checkerboard). The data size is 3891x1210, where each column contains an fMRI slice at a given time instance. The slices are cutting though the visual cortex, oriented in the plane of the calcarine sulcus. (Note that the 3891 pixels is a masked area - the full image slice is 82x68 (5576) pixels). The recording time was 403.3s with a TR = 0.333 s. During this time period, the subject was subjected to ten rounds of stimuli. A fixating cross is present, when the flickering checkerboard is not present.

3.2. Full Volume finger tap paradigm with movement artifacts

The second dataset is part of an fMRI dataset collected at Hvidovre Hospital, (Rasmussen et al., 2012), containing 28 test subjects, instructed to perform a finger tapping paradigm; here we focus on a randomly selected subject. The data is collected with a scan repeat time of TR = 2.49 s over 240 samples (reduced data length). The stimuli changes, alternating between left and right hand, starting with the left hand movement. Each active period is 20 s followed by 9.88 s rest. During the resting period, a fixation cross is shown in the middle of the screen; and in the left and right stimuli conditions there is a visual cue (red/green blinking dot at 1 Hz) to pace movement. The full data volume is $53 \times 63 \times 46$ voxels, but is masked to contain 60678 (out of 153594) voxels holding the brain volume. The preprocessing is kept at a minimum, with simple motion correction, normalization, smoothing and realignment in SPM. No principal component analysis has been applied for data reduction, and no highpass filtering, detrending or spike removal has been applied.

4. Experiments

This section focuses experimental results in order to assess the model performance, and to perform a comparison with standard algorithms. An ICA of the fast acquisition dataset is described in detail in McKeown et al. (2003), and the interpretation of the inferred sources may serve as a reference when comparing to GPICA. We will for bench marking present a

comparison with the arguably most popular ICA algorithm, InfoMax (Bell and Sejnowski, 1995), and furthermore compare GPICA with the Molgedey-Schuster algorithm (MSICA) (Molgedey and Schuster, 1994) to get a comparison to another temporal correlated ICA algorithm. Note that we will only compare GPICA to the temporal decompostion of the other algorithms in order to keep the frame of reference for GPICA. The FastICA (Hyvärinen, 1999) was also tested, but did not yield results better than InfoMax, and have therefore been omitted.

Note that it was decided to perform the analysis on non scandalized data, in order not to amplify low variance noise. It will though require parameter settings to be data specific in order to account for the noise variance level controlled by Ψ. See Section 4.5 for standard settings regarding the supplied toolkit.

In the remainder of the section we discuss the results for the fast acquisition visual paradigm dataset, Section 4.1, and use that as an example to show how to assess convergence of the algorithm, Section 4.2. We will present the convolutive extension, Section 4.3, and finally we discuss the results for the full volume finger tapping data set in Section 4.4. The toolkit application and standard settings on standardized data will be presented in 4.5.

4.1. Fast acquisition visual paradigm

The high sampling frequency (low TR) on the dataset, due to single slice acquisition, implies higher temporal dependency, and hence data very suitable for the GPICA algorithm. A hemodynamic response, such as the BOLD signal, has smooth and often low-frequent content, since the hemodynamic response function (HRF) peaks after approximately 6s (for a dirac delta stimuli), (Martin et al., 2006). In order to accommodate both long and short time scales, the mean of the prior gamma distribution in Eq. 8 is set to match a common timescale, where a mean of 0.4 s (1.2 samples) is chosen. Note that the short TR also allows for observing heart beat related signals, and the prior must therefore also allow for high frequent content. Having no a priori knowledge of the timescales, the algorithm will still converge, but can require more iterations. Note that the hyperparameters for Ψ

is not the most sensitive, as long as a relative flat prior is chosen (Here $s_a = 1, r_a = 1$, confer Section 4.5 for standard settings). For the precision elements, α (of W), the hyper-
205 parameters for the Gamma prior is $s_b = 2, r_b = 1$, and the prior is visualized in Figure B.8 together with the result of the posterior sampling.

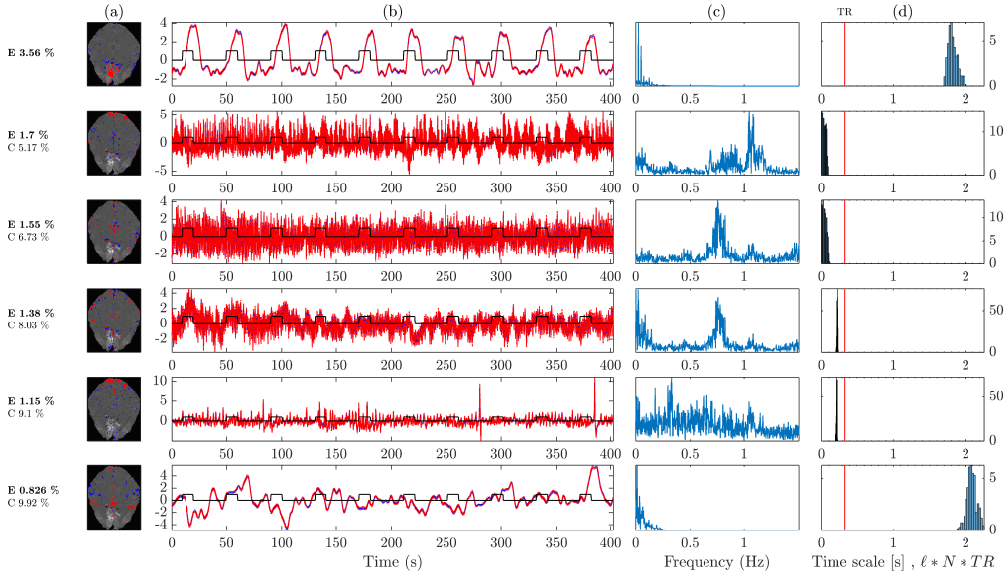
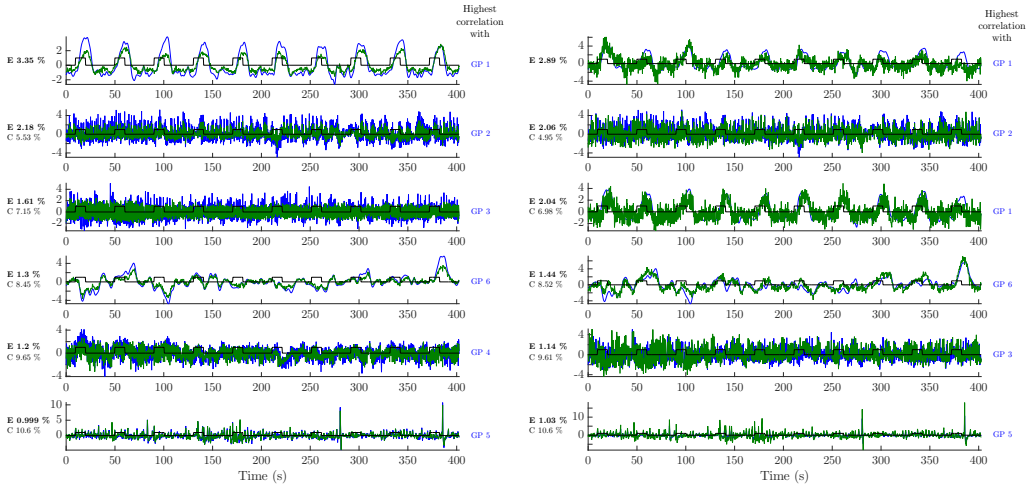


Figure 3: (a) Spatial map of the corresponding time series components in (b). The percentages given to outermost left are the energy by each component and cumulative energy of components included from the top. The definition of energy is given in Section 2.5. The maps (a) show the 2.5 % (blue) and 97.5 % (red) quantiles of the median mixing matrix, superimposed upon an anatomical reference image created by averaging over all acquired images. In (b) the blue median is surrounded by the 5% and 95% quantiles, and the components are sorted according to amount of energy. (c) Frequency spectrum. (d) Marginal distribution of ℓ reported in seconds.

Figure 3 shows the main results for the output of the algorithm. The figure layout follows that of (McKeown et al., 2003). The spatial maps in (a) are generated from the median of each of the rows in W (as described in Section 2.5, and holds the 2.5 % (blue) and 97.5 % (red) quantiles of the median. The time series components in (b), represents the median of each of the rows of Z (blue) and its 5% and 95% quantiles in red. Due to the strong signal in the data and good convergence, the upper and lower quantiles are close to the mean and gives a high reliability in the results. (c) shows the frequency content of the source which is clearly related to the spectra of the marginal distribution of ℓ in (d). *I.e.* the longer timescale of ℓ, the lower the frequency content of the component. Note the red line in (d) that symbolizes the TR. A $\ell < TR$ will still imply prior sample dependency, but with $\ell << TR$ will reduce the algorithm to pPCA. The stimuli related component 1 exhibit a clear paradigm-related activation, and the corresponding spatial map shows a clear activation in the area around the visual cortex. Following the reasoning in (McKeown et al., 2003), component 2 and 4 (more uniform frequency spectre) contains a breath related non-stimuli conponent, whereas component 3 is more likely to capture effects from the heart beat. Component 5 have a broad-band frequency (white noise like) spectrum, and a activation map close to the edge of the brain, which could be a sign of a motion artifact. Component 6 is low frequent, and holds the largest ℓ time scale. Taken together with the spatial map, the component would likely capture vasomotor oscillations (McKeown et al., 2003).

Figure 4 shows the components found by temporal oriented icaMS and temporal Infomax. There is a clear correspondence between all components in icaMS, whereas Infomax is struggling to extract a noise free stimuli holding component. Especially the smoother signal and the concurrently enhanced energy in the GPICA stimuli related component, testifies to a outperformance of both ICAMS and Infomax. Note the lower energy percentages for for both icaMS and Infomax on the stimuli-holding component (3.35 % and 2.04 %, respectively, compared to GPICA with 3.56 %). For comparison reasons we use temporal versions of ICAMS and Infomax, and it should be noted, that the spatial versions of Infomax gives a better stiumli holding representation than the temporal. It furthermore

explains 3.85 % of the energy in the first component, but still struggles to extract smooth noise free activation related time series. Since we aim for a temporal enhancement based algorithm, we choose only to focus on temporal decomposed ICA's.



(a) Temporal icaMS (b) Temporal Infomax

Figure 4: icaMS and Infomax (in green) as a comparison with Figure 3. They are superimposed upon the source from GPICA to whom they have the highest correlation (shown in blue). Note the energy notation in the left hand side.

4.2. Assessing MCMC convergence

The following will give a detailed description on how to asses the the convergence of the GPICA algorithm in general - the individual figure discussion will follow. We will use visual inspection on a selected set of parameters in order to make the interpretation as easy as possible, see Figure 5 for an example. The figure holds the development of the parameter ℓ (leftmost), the acceptance rate (middle) and the loglikelihood (rightmost), all adding an extra information for the overall convergence assessment. Note that all subplots includes the warm-up period to evaluate if a proper warm-up length has been chosen. The development of ℓ summarizes the histogram in Figure 3 (after the warm-up period), and gives an insight to the length scale of the sources. Note that the scale of ℓ in Figure 5 is

reported in normalized time points. High fluctuations in the development of ℓ can testify to a slightly unspecific prior for longer length scales, or contrary when the length scales are so small that they approach the limit for pPCA, where all proposed samples in that limit will be accepted. Furthermore, the first subplot holds the prior of ℓ superimposed in gray, in order to access the support of the prior. An aspect on identifiability can also been assessed by overlapping histograms, or interchanging ℓ developments . The chosen step size (the variance of the proposal distribution) can be evaluated from the sources ability to evolve over iterations, and can moreover be regulated by the development of the acceptance rate in the middle subplot. Extremely high acceptance rates can, as mentioned, be accepted if we are in the limit for pPCA. Acceptance rates should be lower that 60% and optimally approx. 30%. Lastly the loglikelihood is visualized in order to evaluate convergence towards stationarity. Note that all subplots optimally should be stable after the warm-up period.

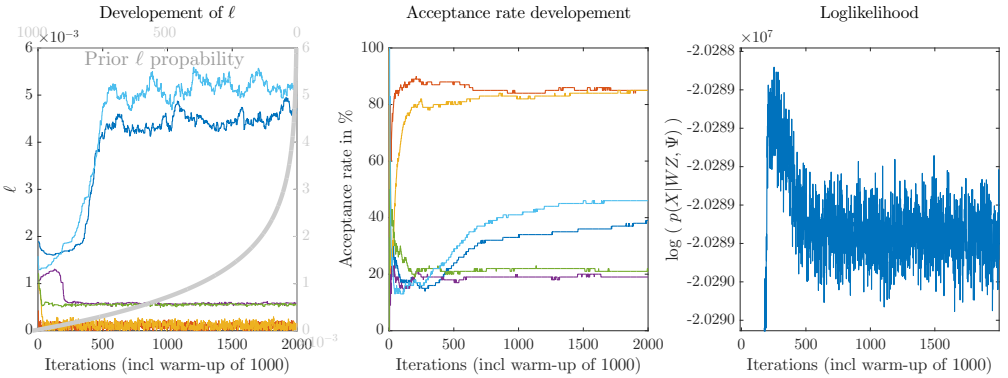


Figure 5: Assessing convergence over function iterations for the data outcome in Figure 3. Following the progressing of ℓ (a), acceptance rate (b) and the loglikelihood (c).

4.3. Convolutive ICA

The stimuliholding component of the convolutive sources are summarized in Figure 6 (all sources in Figure B.9) and Figure B.10. Note that these convolutive results are exploratory and meant for proving the concept of finding time delayed versions of the

sources. In Figure B.9 all investigated lags are visualized with spatial maps and the (zero lag) corresponding time series. It is important to note the clear effect in the spatial maps demonstrating that time delayed versions of the signal do exist, and that they can be captured by convolutive GPICA.

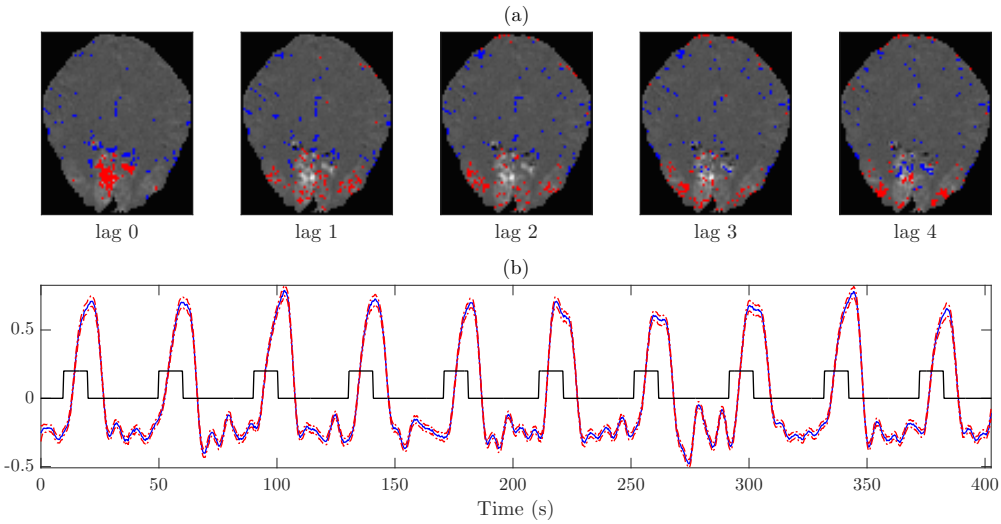


Figure 6: Convolutive main results for one component (compared to instantaneous model in Figure 3). All sources can be seen in Figure B.9 (a) Spatial map for all lags of the corresponding time series components in (b). The maps (a) show the 2.5 % (blue) and 97.5 % (red) quantiles of the mixing matrix's median superimposed upon an anatomical reference image created by averaging over all acquired images. In (b) the blue time source median is surrounded by the 5% and 95% quantiles in red.

4.4. Full Volume data on Finger Tapping Paradigm with Movement Artifacts

In order to demonstrate the methods performance on more standard fMRI data, we applied GPICA on the data from Rasmussen et al. (2012), and the following will outline the results. The lower sampling frequency of the data is expected to force the model to seek for single sample dependency, and the model prior should thus still allow for both high and more low frequent content. Different priors has been explored to facilitate fast convergence, but still maintain the exploratory behavior of the sampler. The ℓ prior is set

based on TR to have a mean around 3s (1.2 samples). Here the hyperparameters for Ψ is $s_a = 1, r_a = 100$, and for the precision elements, α (of W), the hyperparameters for the Gamma prior is $s_b = 2, r_b = 1$,

Figure 7 holds the overview of a 20 component result, and a closeup on the stimuli-holding component (component 19) is shown in B.12. In Figure B.11 the convergence plot similar to Figure 5 is shown for the finger tapping data. Note that all sources actually explains $> 80\%$ of the signal, and that all sources has around one sample dependency, except one with a longer length scale (the stimuli holding component). The first components exhibits a very low frequent nature, and combined with the superficial scalp map, it could be a motion drifting artifact that has not been canceled under the preprocessing. In the 20 component setup, two stimuli holding components are evident (comp. 13 and 19). Comp. 13 appears not to distinguish between left and right hand movement, whereas a comp 19 increase corresponds to a left hand activation, and a decrease to right hand activation (remember the scale ambiguity present).

4.5. Toolkit application

A plug-in for SPM is available from https://github.com/dittehald/GPICA. The plugin serves as a GUI for running GPICA on fMRI data loaded in SPM, and not as a fully integrated part of SPM. The plugin can also be used as a stand alone application, and was created on MATLAB R2015a.

We recommend running an initial run with a low number of samples (*e.g.* 200 warm-up and additional 200 samples), in order to check the initial hyperparameters and settings. Note that the prior for Ψ can be scaled to match an appropriate expected value for both standardized and non-standardized data, confer Eq. A.14. A pre-run is though necessary in order to estaimate the size of the error term. The predefined settings in the plugin are though suitable for most standardized fMRI data. The default settings will set the mean length scale to 1.2 samples based upon the specified TR, but can be customized to fit any expectation. The standard hyperparameter settings are Ψ: $s_a = 1, r_a = 1$, α $s_b = 2, r_b = 1$. The step-size/ variance of the proposal distribution is to be set according ℓ and TR. The

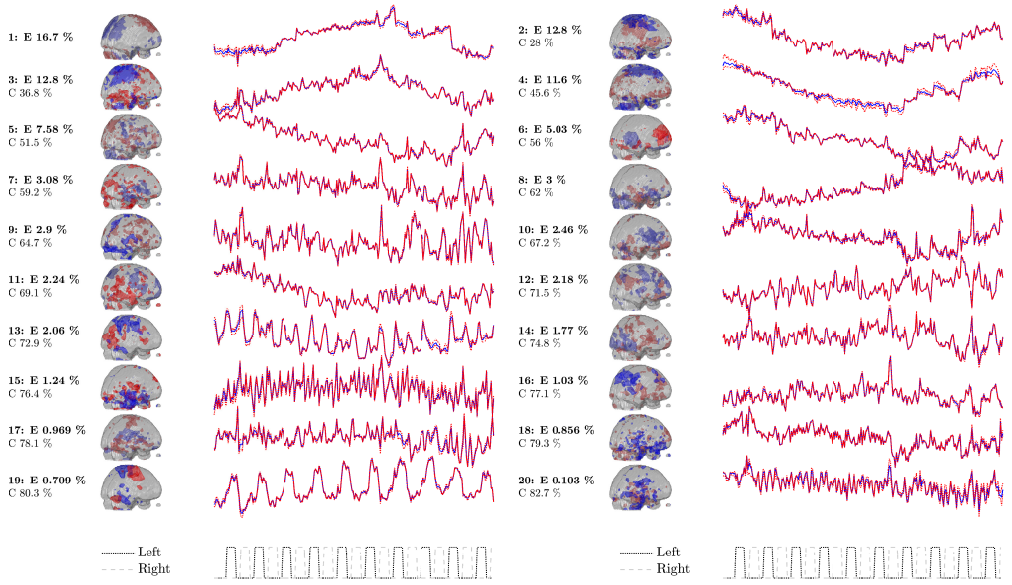


Figure 7: Spatial 3D map and the associated time series components. The maps holds the 2.5 % (blue) and 97.5 % (red) quantiles superimposed upon the anatomical mask. The time series holds the blue median surrounded by the 5% and 95% quantiles, and the components are sorted according to amount of energy.

visual of the convergence assessment will also be available in the plugin, and can guide the user for further optimization.

5. Discussion

We demonstrated the GPICA's potential for fMRI analysis, by analyzing relative unprocessed data from two different datasets (no highpass filtering or detrending, and no prior PCA data reduction). The following will give an insight to the prospect and outlook for the algorithm, as well as the limitations.

For the single slice dataset, we outperform two common used ICA algorithms, by detecting a clear and strong stimuli related component and definable artifacts. The dataset is optimal for GPICA in terms of the short TR and thereby longer sample dependencies. The algorithm is seen to converge after 1000 samples, and the runtime of all 2000 samples was

around 12h on a 8 AMD Quad-Core AMD Opteron(tm) Processor 8356 2.3GHz with 256 GBRAM. As mentioned, the runtime and the model complexity is the major drawback of GPICA. This is usually a reasonable price to pay for getting more discriminative findings. Parallelization and computational optimization could lead to a shorter processing time.

The convolutive exploratory model expansion gives the possibility for extracting additional signal information, and enhances prospective model applications. On the convolutive mixture model, it was possible to extract meaningful mixing matrices, that show a given source process progress through the lags. A dispersion of the source is seen through the lags, and the time line gives a smooth and reliable presentation of the signal. Being able to extract that kind of information may be out to quite important for low TR studies.

The performance on the 3D data testifies to a model capable of detection low energy holding stimuli related components in a highly artifact covered data. Especially the high amount of drift/low frequency oscillation was also detected by the algorithm. Prior high-pass filtering could remove noise and artifacts data, but to support an end-to-end approach to data analysis, it was our purpose to demonstrate the prospect to perform on data with a limited amount of prepossessing. Preproccesing also may introduce an element of over-fitting, for example for this dataset the source signal related to the paradigm is quite weak and easily removed by too aggressive filter. It is not unlikely that the data analyst would stop experimenting with preprocessing settings once a paradigm related source is identified. Therefore keeping the preprocessing to a minimum is in genral a good idea. GPICA also detects clear evaluable areas from the mixing matrix that can give a great insight into the sources locations, so even though GPICA is a temporal based algorithm, the spatial specificity is still keep high.

The proposed method is recommended for time series holding longer timescales in order to facilitate the best utilization of the smoothness of the GPICA. A potential application, that also could take advantages of the a prioi smoothness constrains, would be electroencephalography (EEG). We do often seek a smooth event-related potential ERPs, and for classification, GPICA could be a very promising method. The high sampling rate in EEG setups would facilitate GPICA as a great tool for extracting smoother components, par-

ticularly good for classification of event related potentials (Henriksen, 2012).

It is also possible to initialize ℓ with different more specific prior distributions, but it is not guaranteed, that the different initialized priors always will hold the expected components, due to the random walk of the MCMC. Only a very narrow prior distribution would give higher probability for capturing e.g. pre-known time fluctuations, but it demands a precise a priori knowledge of the dataset. At the moment only temporally independent sources are sought, but it could be extended to spatial-temporal dependency, by expending the approximations, as in (Luttinen and Ilin, 2009).

To summarize, instantaneous and convolutive GPICA can be used as an additional tool in the pipeline for extracting source information in fMRI. It is recommended as an extra tool used *e.g.* after running standard ICA procedures, to gain valuable additional insight in the importance of temporal dependency in the data. GPICA supports end-to-end mentality, as it is not dependent on restricted data preprocessing, that, in worse case, can remove important signal contributions.

6. Conclusion

We have introduced temporal smoothness for the sought source priors, and argued that the perspectives for GP based ICA could take advantage of the temporal dependency seen in neural data. It supports end-to-end methods mentality and can be benchedmarked against well established ICA methods.

7. Acknowledgment

The authors would like to thank DTU for funding the PhD work. A great thanks to Lars Kai Hansen for sharing of knowledge and 2D data and to Kasper Winther Andersen for sharing of the 3D dataset.

AppendixA. Gibbs sampling

All MCMC updates apart from for ℓ which is described in Section 2.3 are handled by Gibbs sampling. The derivation of the conditional distributions and the computational details are given in this appendix. The conditional probabilities are proportional to the joint probability equation (15).

To summarize we have

$$p\left(Z|X,W,\Psi,\ell\right) \propto \mathcal{N}\left(X \mid WZ,\Psi\right)\prod_{p}^{P}\mathcal{N}(z_{[p]}|0,K_{\ell_p}) \tag{A.1}$$

$$p\left(W \mid X,Z,\Psi,\alpha\right) \propto \mathcal{N}\left(X \mid WZ,\Psi\right)\prod_{i}^{D}\prod_{p}^{P}\mathcal{N}\left(w_{ip} \mid 0,\alpha_{ip}^{-1}\right) \tag{A.2}$$

$$p\left(\Psi \mid X,Z,W,s_a,r_a\right) \propto \mathcal{N}\left(X \mid WZ,\Psi\right)\prod_{i}^{D}\Gamma^{-1}\left(\Psi_i \mid s_a,1/r_a\right) \tag{A.3}$$

$$p\left(\alpha \mid W,s_b,r_b\right) \propto \prod_{i}^{D}\prod_{p}^{P}\left[\mathcal{N}\left(w_{ip} \mid 0,\alpha_{ip}^{-1}\right)\Gamma\left(\alpha_{ip}^{-1} \mid s_b,r_b\right)\right] \ . \tag{A.4}$$

It will be convenient in the following to introduce the deviation term:

$$\epsilon = X - WZ \tag{A.5}$$

and the related deviation term including all contributions except from the pth source (Knowles and Ghahramani, 2007, Henao and Winther, 2011):

$$\epsilon_{ij\backslash p} = x_{ij} - \sum_{p'\neq p} w_{ip'}z_{p'j} \ . \tag{A.6}$$

We will also use the $\backslash p$ notation to denote set of stochastic variables where the the pth row/column removed. We exploit the structure of the model to make computationally efficient Gibbs sampling updates as detailed in the following.

AppendixA.1. $p\left(W \mid X,Z,\Psi,\alpha\right)$

For W we sequentially sample source slices $w_p = (w_{1p},\ldots,w_{Dp})^T$ from

$$p\left(w_p \mid X,W_{\backslash p},Z,\Psi,\alpha_p\right) \propto \mathcal{N}\left(X \mid WZ,\Psi\right)\prod_{i}^{D}\mathcal{N}\left(w_{ip} \mid 0,\alpha_{ip}^{-1}\right) \ . \tag{A.7}$$

To isolate the w_p-dependence in the likelihood, we write

$$\mathrm{Tr}\,\Psi^{-1}\epsilon\epsilon^T = \mathrm{Tr}\,\Psi^{-1}(\epsilon_{\setminus p} - w_p z_{[p]}^T)(\epsilon_{\setminus p} - w_p z_{[p]}^T)^T \tag{A.8}$$

$$= z_{[p]}^T z_{[p]} w_p^T \Psi^{-1} w_p - 2 w_p^T \Psi^{-1} \epsilon_{\setminus p} z_{[p]} + \mathrm{Tr}\,\Psi^{-1}\epsilon_{\setminus p}\epsilon_{\setminus p}^T \tag{A.9}$$

We combine this contribution with the corresponding prior term $\sum_i \alpha_{ip} w_{ip}^2$ and by completing the square (identifying mean and variance from matching to $(x-\mu)^T\Sigma^{-1}(x-\mu) = x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu$) (Bishop, 2006) we get

$$p\left(w_p \mid X, W_{\setminus p}, Z, \Psi, \alpha_p\right) = \mathcal{N}\left(w_p \mid \Sigma\Psi^{-1}\epsilon_{\setminus p} z_{[p]}, \Sigma\right) \tag{A.10}$$

with Σ being diagonal with elements $\left(\Psi_{ii}^{-1} z_{[p]}^T z_{[p]} + \alpha_{ip}\right)^{-1}$.

AppendixA.2. $p\left(Z|X, W, \Psi, \alpha, \ell\right)$

We apply the same procedure for Z as for W. The conditional distribution for a time-slice $z_{[p]}$ of Z is proportional to,

$$p\left(z_{[p]}|X, W, Z_{\setminus[p]}, \Psi, \ell\right) \propto \mathcal{N}\left(X|WZ,\right)\mathcal{N}(z_{[p]} \mid 0, K_{\ell_p}) \tag{A.11}$$

As above we complete the square but this time for $z_{[p]}$ to get:

$$p\left(z_{[p]}|X, W, Z_{\setminus[p]}, \Psi, \ell\right) = \mathcal{N}\left(z_{[p]}|\Sigma_z w_p^\top \Psi^{-1}\epsilon_{\setminus p}, \Sigma_z\right) \tag{A.12}$$

with $\Sigma_z^{-1} = w_p^\top \Psi^{-1} w_p I + K_p^{-1}$ and I is the identity matrix. In order to sample from this conditional we need to both compute an $\mathcal{O}(N^3)$ matrix inversion to get Σ_z and to compute the Cholesky decomposition (or matrix square-root) of Σ_z. Rasmussen and Williams (2006) give a numerically stable and fast scheme to achieve both based upon the Cholesky factorization of $K_p + I/w_p^\top \Psi^{-1} w_p$. In our implementation we use their approach.

Hartikainen et al. (2010) showed how to map temporal Gaussian processes to a Kalman filter and thereby achieve $\mathcal{O}(NP^3)$ computational complexity. This in general compares favorable to the $\mathcal{O}(N^3P)$ scheme for sampling from conditionals described above since typically $P \ll N$. We therefore implemented a forward filtering-backward sampling version the Hartikainen et al. (2010) approach. However, we observed substantial computational overhead and some numerical instability so for the length of the time-series considered here we recommend and use the approach described above.

AppendixA.3. Ψ and α

We use conjugate priors for Ψ^{-1} and α so we get standard updates for conditionals:

$$p\left(\Psi^{-1} \mid X, Z, W, s_a, r_a\right) \propto \mathcal{N}\left(X \mid WZ, \Psi\right) \prod_i^D \Gamma\left(\Psi_{ii}^{-1} \mid s_a, r_a\right) \tag{A.13}$$

$$p\left(\Psi^{-1} \mid X, Z, W, s_a, r_a\right) = \prod_i^D \Gamma\left(\Psi_{ii}^{-1} \mid s_a + \frac{N}{2}, r_a + \frac{1}{2}\left[\left(X - WZ\right)\left(X - WZ\right)^\top\right]_{ii}\right) , \tag{A.14}$$

where $[A]_{ij} \equiv A_{ij}$.

For α we get:

$$p\left(\alpha \mid W, s_b, r_b\right) \propto \prod_i^N \prod_p^P \left[\mathcal{N}\left(w_{ip} \mid 0, \alpha_{ip}^{-1}\right) \Gamma\left(\alpha_{ip} \mid s_b, r_b\right)\right] \tag{A.15}$$

$$p\left(\alpha \mid W, s_b, r_b\right) = \prod_i^N \prod_p^P \Gamma\left(\alpha_{ip} \mid s_b + \frac{1}{2}, r_b + \frac{1}{2} w_{ip}^2\right) . \tag{A.16}$$

AppendixB. Supplementary Material

All additional figures not included in the main part will be listed here. Please confer the caption and the main text for further discussion and explanation.

AppendixB.1. Inferred Parameters

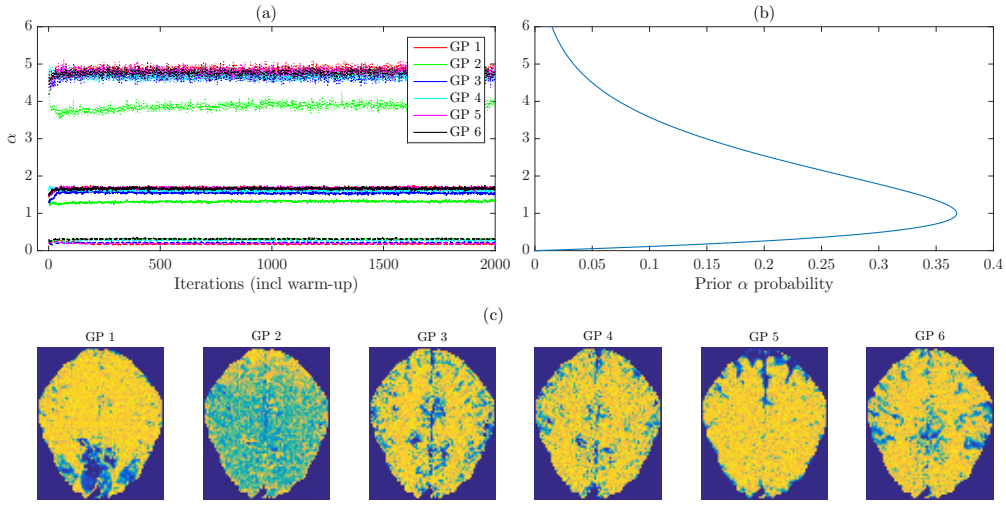


Figure B.8: Visual investigation of α. In (a) the development of α is shown for each component. The median in a full line, and the 2.5% and 97.5% quartiles in dashed. (b) holds the prior for α plotted against the second axis in order to compare to the inferred posterior result in (a). In (c) the median of each component is visualized, and it is clear to see the high variance areas, that corresponds to to the activation in the mixing matrices in 3.

AppendixB.2. Additional Convolutive Results

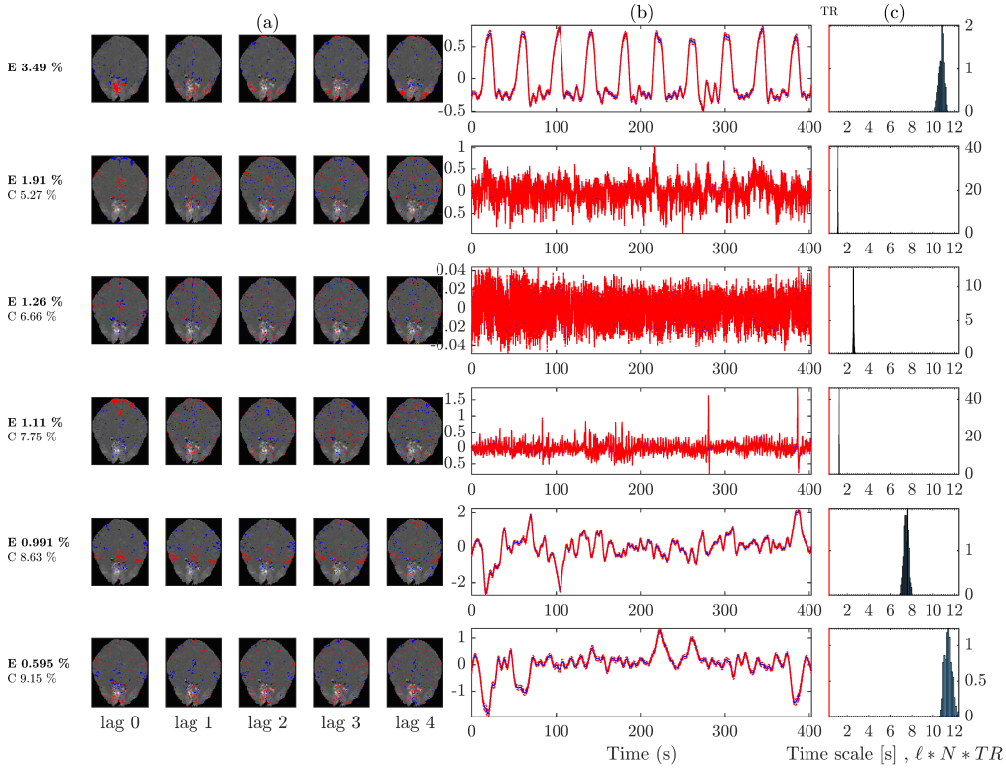


Figure B.9: Convolutive main results (compared to instantaneous model in Figure 3). (a) Spatial map for all lags of the corresponding time series components in (b). The percentages given to outermost left are the energy of each component and cumulative energy of components included from the top. The definition of energy is given in Section 2.5. The maps (a) show the 2.5 % (blue) and 97.5 % (red) quantiles superimposed upon an anatomical reference image created by averaging over all acquired images. In (b) the blue median is surrounded by the 5% and 95% quantiles, and the components are sorted according to amount of energy. (d) Marginal distribution of ℓ measured in seconds.

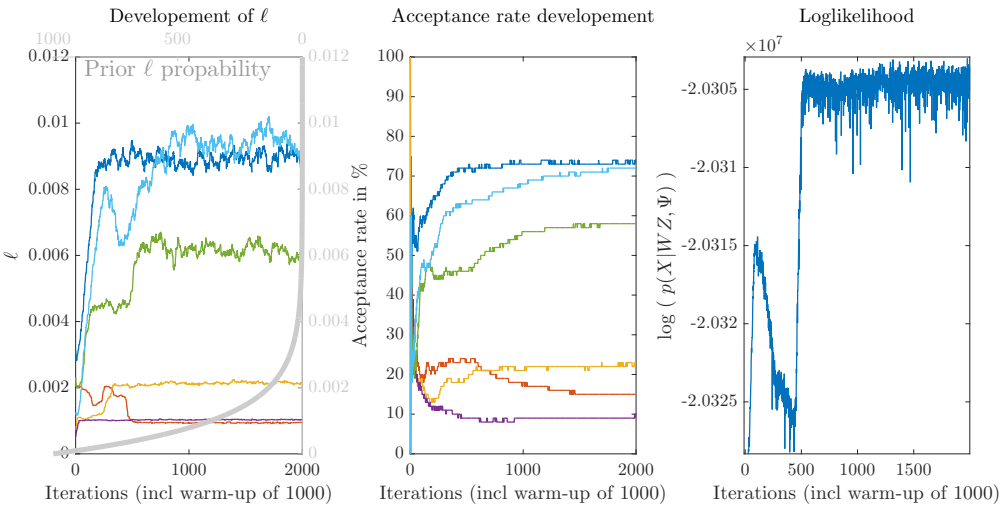


Figure B.10: Assessing convergence over function iterations. Following the progressing of ℓ (a), acceptance rate (b) and the loglikelihood (c). Note the prion of ℓ superimposed upon (a) in gray, and giving a possibility to evaluate the choice of the prior based upon the inferred parameters. All subplots testifies to a converged algorithm.

AppendixB.3. Additional Results on Full Volume Dataset

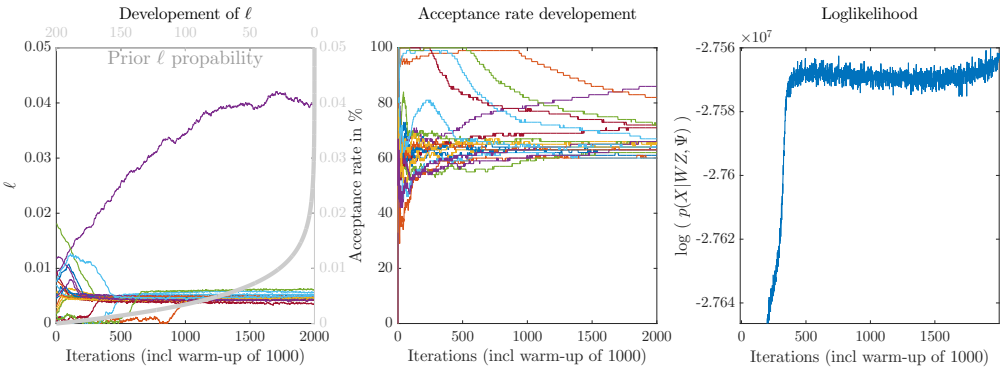


Figure B.11: Convergence diagnostic of Figure 7. The development of ℓ of each sources in the first subplot. Nearly all source components are seen to converge to around a value of 0.005, corresponding to approx. one sample dependency. The single source that demands a longer timescale dependency is the stimuli holding component shown in Figure B.12. The progress of the acceptance rates in subplot three allows for interpretation of the final acceptance rates, and enables control possibilities for the step size. The rightmost subplot holds the log likelihood, where a stabilized log likelihood represents convergences.

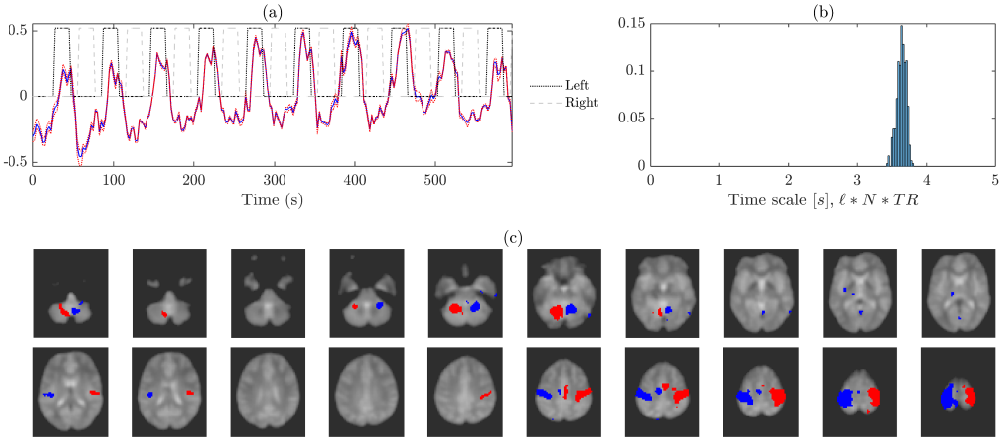


Figure B.12: Component view for the stimuli holding component (capturing both left and right hand movement). Including the histogram over ℓ showing a time dependency of approx. 3.5 s. (c) holds a slice-vice activation map of the mixing matrices. The is a clear correspondence bestrewn the left hand stimuli and the temporal activation of the source. The slice-vice activation maps show a increased activation in the motor cortex corresponding to left hand movement and a smaller decreased activation in the motor cortex for the opposite hand. The crossed cerebellar activation is also present as expected.

AppendixC. References

J.-R. Duann, T.-P. Jung, W.-J. Kuo, T.-C. Yeh, S. Makeig, J.-C. Hsieh, T. J. Sejnowski, Single-trial variability in event-related BOLD signals, Neuroimage 15 (2002) 823–835.

M. J. McKeown, L. K. Hansen, T. J. Sejnowsk, Independent component analysis of functional MRI: what is signal and what is noise?, Curr. Opin. Neurobiol. 13 (2003) 620–629.

A. J. Bell, T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Netw. 10 (1999) 626–634.

A. M. Kagan, C. R. Rao, Y. V. Linnik, Characterization problems in mathematical statistics (1973).

R. Henao, O. Winther, Sparse linear identifiable multivariate modeling, J. Mach. Learn. Res. 12 (2011) 863905.

L. Molgedey, H. G. Schuster, Separation of a mixture of independent signals using time delayed correlations, Phys. Rev. Lett. 72 (1994) 3634–3637.

C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, University Press Group Limited, 2006.

B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, M. Sahani, Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity, J. Neurophysiol. 102 (2009) 614–635.

J. Luttinen, A. Ilin, Variational gaussian-process factor analysis for modeling spatio-temporal data, in: Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22, Curran Associates, Inc., 2009, pp. 1177–1185.

M. N. Schmidt, H. Laurberg, Nonnegative matrix factorization with gaussian process priors, Comput. Intell. Neurosci. (2008) 361705.

M. N. Schmidt, Function factorization using warped gaussian processes, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 921–928.

E. Snelson, C. E. Rasmussen, Z. Ghahramani, Warped gaussian processes, Adv. Neural Inf. Process. Syst. 16 (2004) 337–344.

S. Park, S. Choi, Source separation with gaussian process models, in: Machine Learning: ECML 2007, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 262–273.

S. Park, S. Choi, Gaussian processes for source separation, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pp. 1909–1912.

R. K. Olsson, L. K. Hansen, Blind separation of more sources than sensors in convolutive mixtures, in: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, volume 5, pp. V–V.

J. Hartikainen, H. Jouni, S. Simo, Kalman filtering and smoothing solutions to temporal gaussian process regression models, in: 2010 IEEE International Workshop on Machine Learning for Signal Processing.

T. W. Lee, A. J. Bell, R. Orglmeister, Blind source separation of real world signals, in: Neural Networks,1997., International Conference on, volume 4, ieeexplore.ieee.org, 1997, pp. 2129–2134 vol.4.

L. Parra, C. Spence, B. D. Vries, Convolutive blind source separation based on multiple decorrelation, in: Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop, ieeexplore.ieee.org, 1998, pp. 23–32.

R. K. Olsson, D. K. L. K. Hansen, Linear State-Space models for blind source separation, J. Mach. Learn. Res. 7 (2006) 2585–2602.

M. Dyrholm, L. K. Hansen, CICAAR: Convolutive ICA with an auto-regressive inverse model, in: Independent Component Analysis and Blind Signal Separation, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2004, pp. 594–601.

M. Dyrholm, S. Makeig, L. K. Hansen, Model selection for convolutive ICA with an application to spatiotemporal analysis of EEG, Neural Comput. 19 (2007) 934–955.

L. K. Hansen, ICA if fMRI based on a convolutive mixture model, in: Ninth Annual Meeting of the Organization for Human Brain Mapping,(hbm), New York, June, forskningsdatabasen.dk, 2003.

K. S. Petersen, L. K. Hansen, T. Kolenda, E. Rostrup, S. Strother, On the independent components of functional neuroimages, in: Third international conference on independent component analysis and blind source separation, research.ics.aalto.fi, 2000, pp. 615–620.

P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, S. C. Strother, Model sparsity and brain pattern interpretation of classification models in neuroimaging, Pattern Recognit. 45 (2012) 2085–2100.

C. Martin, J. Martindale, J. Berwick, J. Mayhew, Investigating neural–hemodynamic coupling and the hemodynamic response function in the awake rat, Neuroimage 32 (2006) 33–48.

H. Henriksen, A generative approach to eeg source separation, classification and artifact correction, Master thesis, Technical University of Denmark, 2012.

D. Knowles, Z. Ghahramani, Infinite sparse factor analysis and infinite independent components analysis, in: Independent Component Analysis and Signal Separation, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 381–388.

C. M. Bishop, Pattern Recognition, Springer-Verlag New York, Inc., 2006.

Bibliography

Functional magnetic resonance imaging (fMRI) - NYU cognitive neurophysiology laboratory. https://www.med.nyu.edu/thesenlab/research-0/research-functional-magnetic-resonance-imaging-fmri/. Accessed: 2016-6-15.

ICA:DTU toolbox. http://cogsys.imm.dtu.dk/toolbox/ica/, 2002. Accessed: 2016-6-16.

G K Aguirre, E Zarahn, and M D'esposito. The variability of human, BOLD hemodynamic responses. *Neuroimage*, 8(4):360–369, November 1998.

F Gregory Ashby. *Statistical analysis of fMRI data*. MIT press, 2011.

A J Bell and T J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7(6):1129–1159, November 1995.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magn. Reson. Med.*, 34(4):537–541, 1995.

George E P Box, Norman Richard Draper, and Others. *Empirical model-building and response surfaces*, volume 424. Wiley New York, 1987.

M Brett, W Penny, and S Kiebel. Introduction to random field theory. *Human brain function*, 2004.

E T Bullmore, J Suckling, S Overmeyer, S Rabe-Hesketh, E Taylor, and M J Brammer. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imaging*, 18(1):32–42, January 1999.

Richard B Buxton, Kâmil Uludağ, David J Dubowitz, and Thomas T Liu. Modeling the hemodynamic response to brain activation. *Neuroimage*, 23 Suppl 1:S220–33, 2004.

E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, 1953.

Justin R Chumbley and Karl J Friston. False discovery rate revisited: FDR and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70, 1 January 2009.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

J S Damoiseaux, S A R B Rombouts, F Barkhof, P Scheltens, C J Stam, S M Smith, and C F Beckmann. Consistent resting-state networks across healthy subjects. *Proc. Natl. Acad. Sci. U. S. A.*, 103(37):13848–13853, 12 September 2006.

Arnaud Delorme and Scott Makeig. EEGLAB wiki - SCCN. `http://sccn.ucsd.edu/wiki/EEGLAB`. Accessed: 2016-6-15.

R S J Frackowiak, K J Friston, C Frith, R Dolan, C J Price, S Zeki, J Ashburner, and W D Penny. *Human Brain Function*. Academic Press, 2nd edition, 2003.

John Freund and Richard A Johnson. *Miller and Freund's probability and statistics for engineers*. Prentice Hall, 2011.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

L K Hansen, J Larsen, and T Kolenda. Blind detection of independent dynamic components. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 5, pages 3197–3200 vol.5, 2001.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *Math. Intelligencer*, 27(2):83–85, 2005.

Simon Haykin and Zhe Chen. The cocktail party problem. *Neural Comput.*, 17 (9):1875–1902, September 2005.

G E Hinton and T J Sejnowski. *Unsupervised Learning: Foundations of Neural Computation*. A Bradford Book. MCGRAW HILL BOOK Company, 1999.

A Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, 10(3):626–634, 1999.

A Hyvärinen and E Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13(4-5):411–430, May 2000.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley, 1 June 2001.

Ron Kohavi and Foster Provost. Glossary of terms. *Mach. Learn.*, 30(2-3): 271–274, 1998.

Z J Koles, M S Lazar, and S Z Zhou. Spatial patterns underlying population differences in the background EEG. *Brain Topogr.*, 2(4):275–284, 1990.

Jan Larsen, Lars Kai Hansen, Thomas Kolenda, and Finn Aarup Nielsen. On independent component analysis for multimedia signals. In *in Multimedia Image and Video Processing*, 2000.

D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 21 October 1999.

Jaakko Luttinen and Alexander Ilin. Variational gaussian-process factor analysis for modeling spatio-temporal data. In Y Bengio, D Schuurmans, J D Lafferty, C K I Williams, and A Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1177–1185. Curran Associates, Inc., 2009.

D J C MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Eric Maris and Robert Oostenveld. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods*, 164(1):177–190, 15 August 2007.

Josh H McDermott. The cocktail party problem. *Curr. Biol.*, 19(22):R1024–7, 1 December 2009.

L Molgedey and H G Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634–3637, 6 June 1994.

Morten Møup, Kristoffer Madsen, Anne-Marie Dogonowski, Hartwig Siebner, and Lars K Hansen. Infinite relational modeling of functional connectivity in resting state fMRI. In J D Lafferty, C K I Williams, J Shawe-Taylor, R S Zemel, and A Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1750–1758. Curran Associates, Inc., 2010.

Thomas E Nichols and Andandrew P Holmes. Nonparametric permutation tests for functional neuroimaging: APrimerwith examples. *r H uma n B rai n M appin g*, 1:5–1–25, 2001.

Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.*, 2011:156869, 2011.

Sunho Park and Seungjin Choi. Source separation with gaussian process models. In *Machine Learning: ECML 2007*, Lecture Notes in Computer Science, pages 262–273. Springer Berlin Heidelberg, 17 September 2007.

Sunho Park and Seungjin Choi. Gaussian processes for source separation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1909–1912, March 2008.

Cyril Pernet and Cyril Pernet. Statistical inferencesstatistical inferences in fMRIin fMRI. `http://www.sbirc.ed.ac.uk/cyril/SPM-course/Talks/2013/3-Statistical%20inferences-CP.pdf`, 2016. Accessed: 2016-NA-NA.

M E Raichle, A M MacLeod, A Z Snyder, W J Powers, D A Gusnard, and G L Shulman. A default mode of brain function. *Proc. Natl. Acad. Sci. U. S. A.*, 98(2):676–682, 16 January 2001.

Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. University Press Group Limited, 01 2006.

Peter M Rasmussen, Lars K Hansen, Kristoffer H Madsen, Nathan W Churchill, and Stephen C Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognit.*, 45(6):2085–2100, June 2012.

J H Saltzer, D P Reed, and D D Clark. End-to-end arguments in system design. *ACM Trans. Comput. Syst.*, 2(4):277–288, November 1984.

Mikkel N Schmidt. Function factorization using warped gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 921–928. ACM, 14 June 2009.

Mikkel N Schmidt and Hans Laurberg. Nonnegative matrix factorization with gaussian process priors. *Comput. Intell. Neurosci.*, page 361705, 2008.

Robert Tarjan. Depth-First search and linear graph algorithms. *SIAM J. Comput.*, 1(2):146–160, 1972.

Martijn van den Heuvel, Rene Mandl, and Hilleke Hulshoff Pol. Normalized cut group clustering of resting-state FMRI data. *PLoS One*, 3(4):e2001, 23 April 2008.

Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.*, 102(1):614–635, July 2009.